Detecting patients with breast tumors based on circular RNAs

Zihao Wang
University of Saskatchewan
Saskatoon, Saskatchewan, Canada
jfv472@mail.usask.ca

ABSTRACT

Precise detection of the amount of breast tumor-related circular RNAs (circRNAs) is vital for breast tumor diagnosis and therapy. However, there is little knowledge about circRNA biomarkers related to breast tumors. In this paper, we introduced a method to select 6480 circRNAs that relates to breast tumor among 24,409 circRNAs. To verify the diagnosis performance of selected circRNAs, a 1D convolutional neural network (1D-CNN) model is implemented and input expression data of selected circRNAs to classify people with non-tumor and tumor. The model was trained and tested based on 114 people without tumors and 1135 people with tumors. It achieved an excellent classification accuracy of 93%. This result demonstrates selected circRNAs can be used to diagnose breast tumors.

CCS CONCEPTS

• **Mathematics of computing**; • **Computing methodologies → Machine learning**;

KEYWORDS

Convolutional neural network, Breast tumor, circular RNAs

1 INTRODUCTION

1.1 Motivation

Circular RNAs (circRNAs) are related to RNA molecules that are connected to various diseases, especially in cancer [3]. It's a large class of non-coding RNAs. CircRNAs are formed by a back-splicing process which can form a closed loop. They are more difficult to degrade and more stable than linear RNAs. The amount of circRNAs are enormous in organisms. Moreover, circRNAs not only can regulate gene expression because of the potential to act as miRNA sponges or decoys but also have more functions than the host gene due to the longer half-life linear RNAs [5].

However, a precise function is lacking for the large majority of circRNAs [3]. This study proposed a method to select circRNAs that are related to breast tumor diagnosis. Then, the detection performance of selected circRNAs is tested by a 1D-CNN model.

1.2 Related works

The role of circRNAs in breast tumors is unclear. Based on their characteristic in sponging disease-specific miRNAs, the circular RNAs in breast cancer was analyzed in the wet lab to find top-ranked circRNAs based on their performance of expression in breast tumor cells and other cancer tissues [1]. The result shows the circRNA hsa_circ_001783 was significantly correlated to breast tumors. Gene Ontology and matrix factorization methods are also able to identify breast tumor-related circRNAs [4].

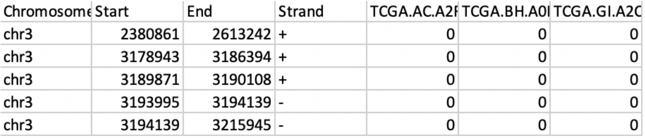
Some studies have used machine learning models to predict cancer types by using RNA gene expression data [2]. Some CNN models were implemented to classify non-tumor and tumor samples.

2 METHODOLOGY

2.1 Dataset and preprocessing

The dataset is provided by Professor Dr. Fangxiang Wu from the biomedical engineering department at the Usask. There are 4 files in the dataset. The first and the second file are people profile data represented by 54,746 circRNAs. The first one includes 114 people without tumors; the second one has 1142 people with tumors. The third one gives the criteria for merging small circRNAs into large circRNA. The fourth one gives 66 circRNAs that are related to breast tumors.

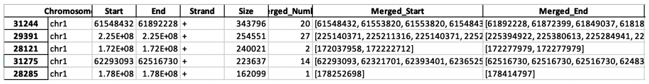
In the people profile data, the useful data fields are: 1) Chromosome: describes which chromosome that a given circRNA is located in. 2): Start: the start point of a given circRNA. 3) End: the endpoint of a given circRNA, followed by 114 (1142) columns – person's name. The example of the people profiles data is shown in Figure 1. For each person, the number means the amount of the corresponding circRNA in a sample.



Chromosome	Start	End	Strand	TCGA.AC.A2F	TCGA.BH.A0I	TCGA.GI.A2C
chr3	2380861	2613242	+	0	0	0
chr3	3178943	3186394	+	0	0	0
chr3	3189871	3190108	+	0	0	0
chr3	3193995	3194139	-	0	0	0
chr3	3194139	3215945	-	0	0	0

Figure 1: The example of the people profiles data

The third file contains criteria for merging small circRNAs into larger ones. The useful data fields are: 1) Chromosome: describes which chromosome that a given circRNA is located in. 2) Start: the start point of a merged circRNA. 3) End: the endpoint of a merged circRNA. 4) Merged_Start: the set of start points of small circRNAs. 5) Merged_End: the set of endpoints of small circRNAs. The example of merging criteria data is shown in Figure 2.



	Chromosom	Start	End	Strand	Size	rged_Num	Merged_Start	Merged_End
31244	chr1	61548432	61892228	+	343796	20	[61548432, 61553820, 61553820, 6154843	[61892228, 61872399, 61849037, 61818
29391	chr1	2.25E+08	2.25E+08	+	254551	27	[225140371, 225211316, 225140371, 2252	[225394922, 225380613, 225284941, 22
28121	chr1	1.72E+08	1.72E+08	+	240021	2	[172037958, 172222712]	[172277979, 172277979]
31275	chr1	62293093	62516730	+	223637	14	[62293093, 62321701, 62393401, 6236525	[62516730, 62516730, 62516730, 62483
28285	chr1	1.78E+08	1.78E+08	+	162099	1	[178252698]	[178414797]

Figure 2: The example of merging criteria data

The first step of preprocessing is merging 54,746 circRNAs into 24,409 circRNAs. Then, removing people who didn't detect any cricRNAs, and removing circRNAs that neither be detected in non-tumor nor tumor groups of people. Finally, there are 114 people in the non-tumor group and 1135 people in the tumor group.

2.2 Selection of breast tumor-related circRNAs

2.2.1 CircRNA frequency. The circRNA frequency, f_i, is used to describe the percentage of people who have the given circRNA. It's come from another concept: genotype frequency. Genotype frequency in a population is the number of individuals with a given genotype divided by the total number of individuals in the population. In this study, we use circRNA frequency to describe the proportion of circRNAs instead of genotype. The definition is described as follows.

The vector $\vec{x}_i$ represents the amount of the i^th circle RNA in each person. The binary classification vector $\vec{y}_i$ is used to describe the value of each element in $\vec{x}_i$ belongs to zero or non-zero. The calculation of $\vec{y}_i$ is as follows:

$$\vec{y}_i[j] = \begin{cases} 0 \ (if \ vecx_i[j] = 0) \\ 1 \ (if \ vecx_i[j] \neq 0) \end{cases} \tag{1}$$

And the circRNA frequency is calculated as:

$$f_j = \frac{\sum_{j=1}^{n} \vec{y}_i[j]}{n} \tag{2}$$

2.2.2 The distance in the circRNA frequency between the 2 groups. For each circle RNA i, the distance of the circRNA frequency (dis_i) is defined as:

$$dis_i = f1_i - f2_i \tag{3}$$

The $f1_i$ means the circRNA frequency for circRNA i in the group without tumor, vice versa. For each chromosome, the average distance of the circRNA frequency is calculated separately based on:1) whole distance of the circRNA frequency, the positive value of the distance of the circRNA frequency, and the negative value of the distance of the circRNA frequency. The average value for chromosome 1 is illustrated in Figure 3. The red line is the average value among the whole distance of the circRNA frequency. The green line is the average value of the positive value. The yellow line is the average value of the negative value.

In each chromosome, circRNAs with distance above or equal to the green line and below the or equal to yellow line are selected. There are 6448 selected circRNAs among all chromosomes. After getting the union between selected circRNAs and the known circRNAs, 6480 circRNAs are used to train the 1D-CNN model.

2.3 1D-CNN with vectorized input

For each person, a length of 6480 vector is the input of the model. In the vector, each element is the amount of the corresponding circRNA in a detected sample. The one-dimensional kernels are applied to the input vector. The output of the 1-D convolutional layer with 32 kernels is passed to a maxpooling layer, a fully connected layer (128 neurons), and a prediction layer (2 neurons). The stride of kernels is equal to the kernel size. In this study, the kernel size is 81. The kernel size of maxpooling is 2. The structure of the CNN model is shown in Figure 4.

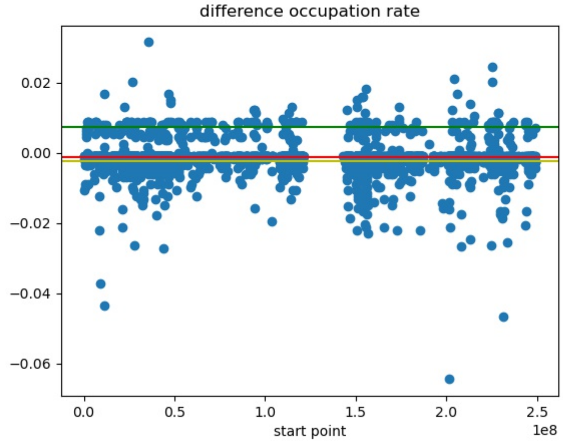


Figure 3: Distance of circRNA frequency in chromosome 1

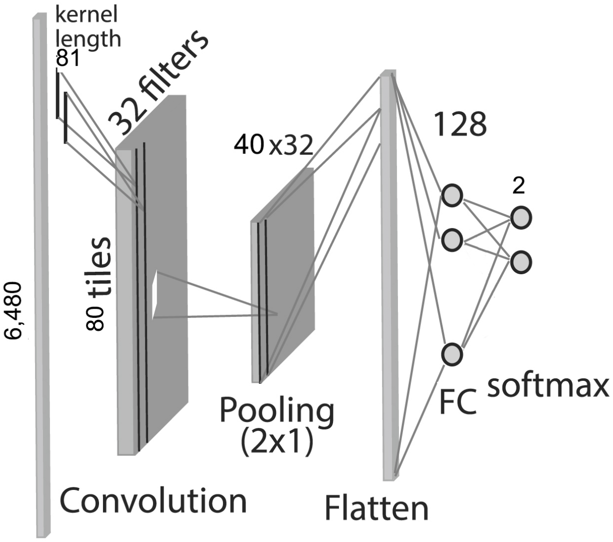


Figure 4: The structure of the CNN model

In addition, categorical cross-entropy, adam optimizer, and the learning rate of 0.001 is selected for the model. The 5-fold cross-validation is applied by choosing 50 epochs and 128 batch sizes for each fold. Respectively, if the training accuracy didn't change in a continuous 3 epochs in each fold, the early stopping is stopped at the third continue epoch in the current fold. Finally, ReLu was used as the activation function and softmax as the prediction layer.

The categorical cross-entropy (CE) is defined as:

$$CE(\vec{x}) = -\sum_{i=1}^{C} y_i * log f_i(\vec{x}) \tag{4}$$

x means the input vector, C is the number of classes. y_i is the actual label in a one-hot vector and $f_i(x)$ is the predicted value in the output vector.

The ReLu activation function of each neuron is:

$$ReLu(x) = \begin{cases} x \ (if \ x > 0) \\ 0 \ (if \ x \leq 0) \end{cases} \tag{5}$$

The softmax function is:

$$softmax(\vec{x})_i = \frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}} \tag{6}$$

In the non-tumor group, 100 people are randomly selected into the training dataset; the remaining 14 people are put into the testing set. In the tumor group, the sample size for training and testing datasets are 1000 and 135 by random selection. Due to the unbalance of the training dataset, the data argumentation is applied by duplicate non-tumor training data 9 times. For each input vector, a non-zero element is randomly selected and added one in each duplication. Finally, training and testing data sizes are 2000 and 149.

3 RESULTS

3.1 The evaluation of circRNA selection

66 known circRNAs are related to the breast tumor, marked as set A. The set of selected circRNAs calls set B. The recall is defined as:

$$recall = \frac{|A \cap B|}{|A|} \tag{7}$$

The selection based on the mean value of the distance of the circRNA frequency in each chromosome is compared with the Chi-square test and Z-test. The comparison of the 3 methods is shown in Table 1.

Table 1: Methods comparison

	p-value	recall	Number of selected circRNAs
Chi-square	<0.05	0.015	206
CHi-square	<0.5	0.23	3236
Z-test	<0.05	0.06	1471
Z-test	<0.5	0.52	5460
distance of circRNA frequency		0.52	6446

When the p-value is less than 0.05, the recall of both the Chi-square test and the Z-test is low. The threshold of p-value 0.5 is also considered. When the Z-test p-value is less than 0.5, the intersection of the selected circRNAs with the known tumor-related circRNAs is as same as the distance of the circRNA frequency method. Based on the best recall and the maximum number of selected circRNAs, the method of setting the mean value for the distance of circRNA frequency gets the best performance.

3.2 Tumor diagnose

There are 19 epochs in the entire training process. Each fold had no more than 50 epochs due to the early stopping manipulation. The training loss and the validation loss are shown in Figure 5. The train and validation dramatically decrease within the first 2 epochs and coverage to 0 after the 7^{th} epoch.

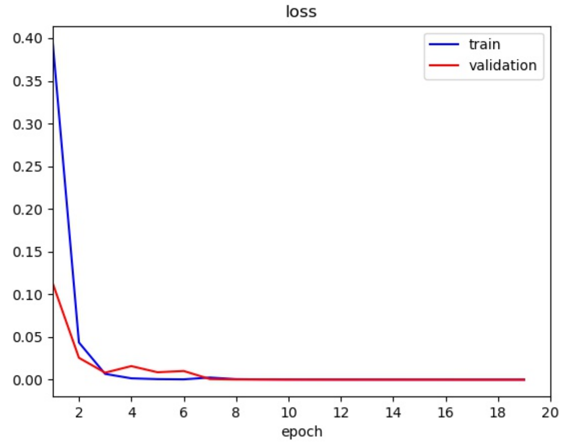


Figure 5: Training and validation loss

The accuracy and the loss for the testing set are 93% and 0.4. The confusion matrix is shown in Table 2. The number shows the actual group and predicted group. In the non-tumor group, 5 out of 14 people are successfully recognized normal samples. In the tumor group, 134 out of 135 people are detected breast tumors.

Table 2: The confusion matrix of the testing set

predict label	Non-tumor	Tumor
Non-tumor	5	9
Tumor	1	134

3.3 Discussion

The curve of training and validation shows the learning process is effective. Moreover, the accuracy shows the 1D-CNN can get effective performance in breast tumor diagnosis. Furthermore, the confusion matrix shows the model can distinguish people with tumors and without tumors.

4 CONCLUSIONS

The purpose of this study is to find circRNAs that can diagnose breast tumors. For each chromosome, the average distance of circRNA frequency was calculated separately based on the positive and negative values. Next, circRNAs with the distance of circRNA frequency above and below the 2 average values are selected. Then, the union between the set of selected circRNAs and the known breast tumor-related circRNAs is defined as the final selection features. These features were input into a 1D-CNN to show if they are related to breast tumors by performing a binary classification task. The high accuracy demonstrated the final selected circRNAs are effective in breast tumor detection. In the future study, the relationship between selected circRNAs and the breast tumor will be ranked to propose effective therapy methods.

REFERENCES

[1] Zihao Liu, You Zhou, Gehao Liang, Yun Ling, Weige Tan, Luyuan Tan, Robert Andrews, Wenjing Zhong, Xuanxuan Zhang, Erwei Song, and Chang Gong. 2019. Circular RNA hsa_circ_001783 regulates breast cancer progression via sponging miR-200c-3p. *Cell Death & Disease* 10, 2 (Feb. 2019), 55. https://doi.org/10.1038/s41419-018-1287-1

[2] Milad Mostavi, Yu-Chiao Chiu, Yufei Huang, and Yidong Chen. 2020. Convolutional neural network models for cancer type prediction based on gene expression. *BMC Medical Genomics* 13, S5 (April 2020), 44. https://doi.org/10.1186/s12920-020-0677-2

[3] Marieke Vromman, Jo Vandesompele, and Pieter-Jan Volders. 2021. Closing the circle: current state and perspectives of circular RNA databases. *Briefings in Bioinformatics* 22, 1 (Jan. 2021), 288–297. https://doi.org/10.1093/bib/bbz175

[4] Shuyuan Wang, Peng Xia, Li Zhang, Lei Yu, Hui Liu, Qianqian Meng, Siyao Liu, Jie Li, Qian Song, Jie Wu, Weida Wang, Lei Yang, Yun Xiao, and Chaohan Xu. 2019. Systematical Identification of Breast Cancer-Related Circular RNA Modules for Deciphering circRNA Functions Based on the Non-Negative Matrix Factorization Algorithm. *International Journal of Molecular Sciences* 20, 4 (Feb. 2019), 919. https://doi.org/10.3390/ijms20040919

[5] Cheng Yan, Jianxin Wang, and Fang-Xiang Wu. 2018. DWNN-RLS: regularized least squares method for predicting circRNA-disease associations. *BMC Bioinformatics* 19, S19 (Dec. 2018), 520. https://doi.org/10.1186/s12859-018-2522-6