INTERNATIONAL JOURNAL OF BIOLOGY AND BIOMEDICAL ENGINEERING
DOI: 10.46300/91011.2022.16.17 Volume 16, 2022

Data analysis for microRNA and related diagnoses

Eugenia Namiot[1], Maxim Khakhin[2]
[1]I.M. Sechenov First Moscow State Medical University (Sechenov University), 125007 Moscow, Russia,
[2]Moscow Aviation Institute, 125993 Moscow, Russia

Received: June 3, 2021. Revised: November 18, 2021. Accepted: December 22, 2021. Published: January 5, 2022.

Abstract— MicroRNAs are non-coding molecules that play a significant role in the development of the disease. MicroRNAs can act as biomarkers or independently lead to the development of a disease. Due to the large numbers of microRNAs, most of the current works focus on the creation of a new way of microRNA clustering or grouping. Today, there are a huge number of different databases that distribute open microRNAs into groups. The problem is that there is no way to evaluate such databases and created clusters. In this work, we propose a new method for assessing the distribution of microRNAs in a cluster, which in the future can be used to predict new sequential ones capable of causing disease. The proposed method can also be used for a better understanding of the mechanisms of various diseases. Since cardiovascular diseases rank first in terms of the number of deaths, they were chosen as the analyzed ones. The Human microRNA Disease Database was used as an analyzed database in this work. The obtained results show that the proposed method can analyze the created databases and can be used in further practice. The proposed model makes it possible to predict new microRNAs for given diagnoses.

Keywords— microRNA, database, cluster analysis, cardiovascular diseases, validation model.

I. Introduction

MicroRNAs are non-coding molecules involved in the regulation of gene expression, most often by binding to mRNA. In addition to the main mechanism, which consists in the directed degradation of mRNA due to the formation of complexes and further inhibition of translation, there are data indicating the ability of microRNAs to participate in methylation processes [1]. A seed region is a certain segment of the mature microRNA sequence mutations of which are often associated with various hereditary diseases. Various microRNA molecules appear as a biomarker of the developed disease. It is assumed that microRNAs similar to each other will cause similar or the same diseases [2]. This is the basis of many studies, when the similarity of microRNAs is determined in various ways, to theoretically predict the possibility of using microRNAs as a biomarker of diseases. The small size of microRNAs and the possibility of their formal representation as a chain of nucleotides with a fixed base (dictionary) make them a popular object for the use of various methods of data analysis.

Cardiovascular diseases (CVD) are the leading cause of death in the world, therefore, various diagnostic methods are being actively developed, the search for biomarkers, as well as a more detailed study of pathogenesis. Despite the importance of the problem, to date, only a small part of all microRNAs involved in the development of CVD has been discovered. There are numerous experimental works describing the role of microRNA in the development of such pathologies: coronary artery disease, hypertrophic changes in the heart chambers, hypertension, including hypertension of the pulmonary vessels, as well as various arrhythmias and myocardial infarction. Moreover, some microRNAs are capable of serving as biomarkers of heart muscle damage, complementing the existing ones [3].

There is a need to find a way to quickly assess the created classification of microRNAs and to predict possible sequences of other microRNAs that can cause diseases.

As a solution to the problem of classifying microRNAs, a large number of databases have been created. Not all of them got a classification based precisely on the assignment of various microRNAs to a specific diagnosis. This article focuses specifically on databases that contain bundles microRNA - the diagnosis of the disease.

The purpose of this work is to study such classifications of microRNAs. The authors propose a new model for assessing such classifications, which can be used both to assess the

quality of existing microRNA groupings and to predict new strands relevant to the diagnosis under study. In our opinion, the described model and the results obtained are new; we have not found any analogues of this development.

The rest of the article is structured as follows. Section 2 describes similar work. Section 3 presents databases with the clinical grouping of microRNAs. Section 4 describes the developed model and the performed computational experiments. Section 5 is devoted to discussion and possible directions of development. Section 6 provides a conclusion.

II. On related works

MicroRNAs can be considered biomarkers for various diseases. The main problem in finding biomarkers using machine learning methods is the difficulty of defining selection criteria. Due to a large number of microRNAs, the analysis of huge datasets, where each microRNA is evaluated separately from all the others, takes a lot of time and often does not show sufficient efficiency. In one of the studies, it was proposed to pre-cluster microRNAs to then compare only the most representative ones among the cluster. After removing uninformative microRNAs, based on the Welch t-test, the authors used hierarchical clustering before proceeding to the main stage of the study [4]. In general, clustering is one of the most commonly used approaches in biomedical research [17].

MicroRNAs can also be considered therapeutic targets. It is necessary to analyze both changes in the regulation of microRNA and the difference in the associations of microRNA-target in the same disease. The work of Tran N. et al. Proposes a new way of microRNA clustering according to similar target groups [5].

Moreover, hierarchical clustering makes it possible to assess the role of microRNAs in different variants of the development of the same disease [6]. By clustering microRNAs, one can also determine whether there are similarities between microRNAs that cause various diseases [7].

Clustering is used to identify microRNA regulatory modules. The identification of such regulatory sites will allow a more detailed study of the mechanism of regulation and interaction of microRNAs during the development of the disease [8].

Clustering methods are used not only to create new ways to study microRNA interactions but also to conveniently present information on specific diseases. The work indicated that there is a fairly strong connection between microRNAs in one cluster and their biological action [9].

Clustering is also used to analyze the processing of microRNAs. One of the studies was based on the fact that the nuclear processing of microRNA is carried out using a specific microprocessor consisting of several proteins. Clustering in this case is used to combine certain microRNAs that have a similar affinity for the microprocessor [10].

The main conclusion from the analyzed literature is that microRNA clustering is mainly aimed at creating new ways of classifying already discovered microRNAs. In turn, no work was found to evaluate the existing classifications, such as checking the accuracy of a specific database with associative links of microRNA - diagnosis.

III. MicroRNA databases

Upon the request "microRNA database" a large number of existing databases are showed and each of them has an individual structure and way of sequence classification. Almost all databases have a common principle of filling with new microRNAs. They are based on experimental or clinical work, where the sequence has been received.

The most frequently mentioned database is miRBase Sequence Database, the main aim of which is to collect sequences of discovered microRNAs [11]. By choosing a specific name for an organism the database produces a complete list of microRNAs related to that organism.

The Comparative Genomic microRNA database (CoGemiR) is another prime example of the frequently mentioned databases. The CoGemiR base is a publicly available database designed to show the genomic organization of microRNAs [12].

MicroRNA Target Prediction database (MiRDB) is one of the new databases created to predict the target of microRNAs. This database also provides a more detailed functional description of microRNAs. Due to the implementation of the MirTarget algorithm, MiRDB can work with new user-defined sequences, selecting a possible target for each of them [13].

However, the databases listed above are mainly used by biologists and engineers, while clinicians are more interested in the relationship between microRNAs and diagnoses. From a formal point of view, such databases describe mappings between sets of microRNA and diagnoses:

$$\{microRNA_1, microRNA_2, \ldots microRNA_n\} \rightarrow Disease_i \quad (1)$$

The purpose of this work is to analyze the quality of such mappings.

As an example of a database that distributes microRNAs depending on the diagnosis, miRCancer can be cited. MiRCancer uses a text-mining algorithm that allows the isolation of various microRNAs that can cause different forms of oncology [14]. Another example is the Mir2Disease database [15].

In our case, the Human microRNA Disease Database (HMDD - https://www.cuilab.cn/hmdd) turned out to be the most convenient for searching of microRNAs related to specific diagnosis. This database contains 35547 microRNA-

diagnosis bundles, 893 diagnoses collected from 19280 articles [16].

IV. On the Model and Performed Computational Experiments

The proposed model is based on the following assumption. If in the diagnostic classifier several microRNAs are assigned to the same diagnosis, then these chains should be similar. Unlike existing works, we are not trying to classify microRNAs. We accept the existing disease classification as a cluster. All microRNAs related to a specific disease (1) are considered as elements of one cluster. The aim of this work is to study the homogeneity of this cluster. We assume that microRNAs assigned to the same diagnosis should be located tightly to each other in the selected space. Violation of such a dense arrangement is the goal of our work.

A. On data embedding

We used the representation of each microRNA as a vector with a frequency of k-mer. Based on the short length of each microRNA, 3-measures were chosen as a characteristic. Since there can only be 4 different nitrogenous bases in any sequence (U, C, G, A), there are 64 possible 3-mers for each sequence. Accordingly, each sequence was encoded with a vector of length 64, where each element corresponded to the number of occurrences of a certain 3-mer in this sequence. In this case, all microRNAs are represented the same, which allows us to calculate the Euclidean distance between them in 64-dimensional space.

B. On the algorithm

In the selected cluster (microRNAs assigned to a specific diagnosis), its diameter (maximum distance between a pair of microRNAs) is calculated. For any disease related cluster C diameter D is $\max_{i,j\in C}\{d(i,j)\}$

Next, we try to exclude from the cluster the chains corresponding to such maximally distant microRNAs and evaluate the change (decrease) in the diameter. A significant (empirically) decrease indicates the remoteness of the selected strands (strands) from the rest of the microRNA group of the cluster, while a slight decrease indicates a dense cluster (close arrangement of the remaining elements). It is the deleted elements that are of interest. Possible options are:

- dense arrangement of microRNAs in the cluster (exclusion of microRNAs slightly reduces the diameter)
- there is the only microRNA, the exclusion of which significantly changes the diameter
- there are several microRNAs, the exclusion of which significantly changes the diameter

The presence of microRNAs in a cluster that significantly affects the diameter may mean the following:

a) error in the database classifier

b) the presence of some microRNAs unknown to us, which should have been in this cluster "between" the main group and "remote" elements. The possibility of predicting such chains is discussed below.

c) the presence of sub-clusters in the original classifier. This may be due to the fact that microRNAs initially assigned to the same diagnosis may have different effects (different mechanisms of action) in relation to a given disease. This can be used to determine further directions of experimental research. It is illustrated in Fig. 1.

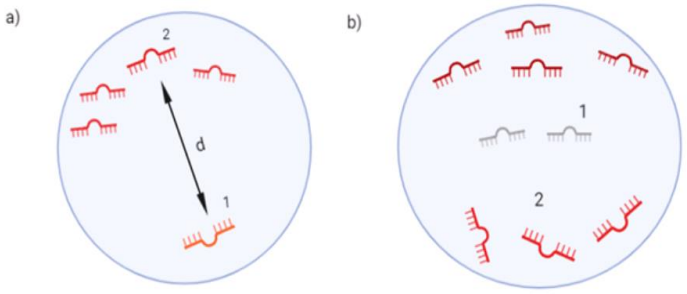

Figure 1. Different versions of the results when analyzing clusters. a - only one microRNA is distant from the general group of sequences, most likely an error in the database classifier (1 - spaced apart from the general group of microRNAs, 2 - densely located microRNAs, d - diameter) b - the presence of some microRNAs unknown to us, which should have been in this cluster "between" the main group and "remote" elements (subcluster). (1 - unknown microRNAs, 2 - subcluster)

V. On the Results

MicroRNAs causing cardiomegaly, coronary artery disease, as well as a generalized group of sequences causing CVD - Cardiovascular diseases (unspecific) were copied from the HMDD database as study groups. Technically, this could be done for any diagnosis for which there are more than three microRNAs.

In total, 9 sequences were isolated for coronary heart disease, between which the pairwise distance was measured. Each of the 9 sequences was considered as a vector of 64 possible 3-measures. The maximum distance (cluster diameter) was between hsa-mir-206 and hsa-mir-31 and equaled 8.4853. Figure 2 shows an initial histogram of the distribution of distances between different sequences for a given diagnosis.

Then, we successively removed one of the sequences included in the pair, which makes up the diameter, until the decrease in the diameter stopped (Fig. 3).

Thus, a sub-cluster of three elements (hsa-mir-206, hsa-mir-31, and hsa-mir-361) was identified in the indicated diagnosis.

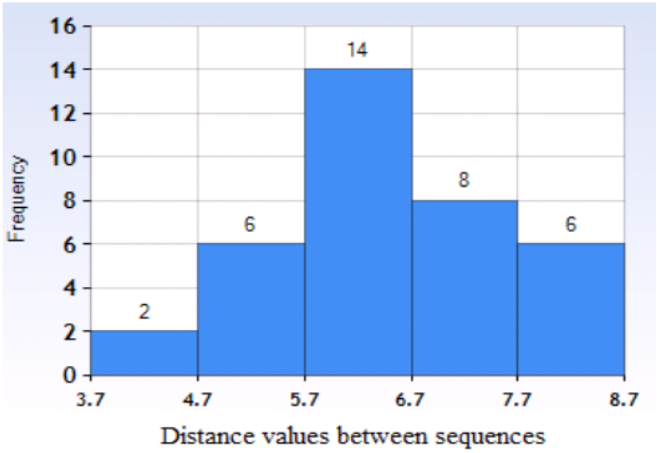

Figure 2. Histogram of the distribution of distances between microRNA sequences involved in the development of coronary heart disease.

Calculations with two other groups of diseases were carried out in a similar way. Thus, in the group of microRNAs involved in the development of cardiomegaly, only one sequence, which is separated from the others, was found, more precisely, has-mir-214. With further successive removal of the sequences, the diameter changed by no more than 2%. This indicates that the remaining sequences are tightly grouped.

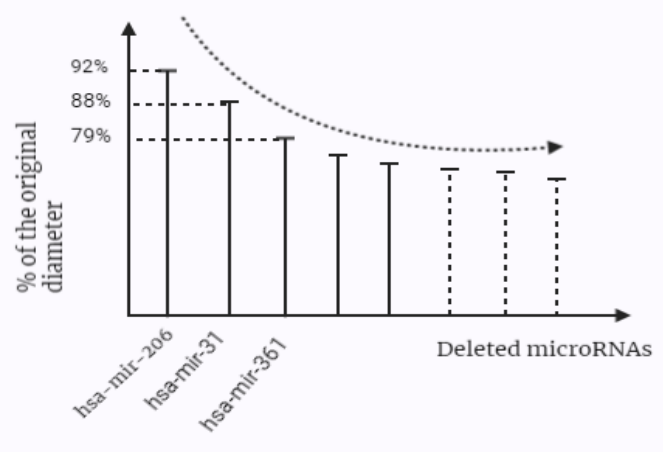

Figure. 3 Relative reduction in diameter

The group with nonspecific cardiovascular diseases was the most numerous; however, when the sequences were removed, the diameter changed insignificantly. At the last step, it changed by only 1% compared to the previous one. Consequently, this group was initially densely formed and there are no spaced sequences in it.

In the group of microRNAs involved in the development of cardiomegaly, only one sequence hsa-mir214 is noticeably separated from all others. Since it is the only sequence, one of the possible options is the initial error in the formation of the cluster. Because of the rather small amount of sequences in the cardiomegaly group, another variant is possible, in which a strong difference in hsa-mir-214 indicates the presence of other microRNAs capable of causing cardiomegaly.

The function of hsa-mir-214 is to develop cardiac hypertrophy and further dysfunction of the heart muscle. Interestingly, there is another microRNA that can stimulate the development of hypertrophy. The role of hsa-mir-206 was to protect the heart muscle from ischemia and reperfusion injury by developing hypertrophy. The distance between hsa-mir-206 and hsa-mir-1 was the smallest in the entire cluster, which indicates a strong similarity of these microRNAs. In this case, hsa-mir-1 has a completely opposite function, which is to prevent the development of cardiomegaly. Such a strong similarity between hsa-mir-206 and hsa-mir-1 indicates a high likelihood of a similar mechanism of action; therefore, further experimental research is needed to clarify the function of both microRNAs in the development of cardiomegaly.

In the group of microRNAs that cause coronary heart disease, three sequences were noticeably different from all others: hsa-mir-206, hsa-mir-31 and hsa-mir-361. One of the possible reasons for the presence of such a subcluster of three microRNAs is the different functions and mechanisms of disease development in microRNAs in different subclusters. The most obvious division of microRNAs in the group was their division according to the final effect on the state of the coronary vessels as a result of the increased expression. However, among the two subclusters, both microRNAs that alleviate the disease and microRNAs that lead to more severe disease were found.

The presence of three different microRNAs may also indicate an incorrectly formed cluster; however, the mechanism by which a denser cluster should be formed is not entirely clear. It is possible that an additional study of these microRNAs is necessary to detect the presence of any common characteristics that differ from the characteristics of other sequences densely located in the cluster.

Possibly, the reason for the heterogeneity of the cluster may lie in the experimental studies themselves, in the course of which the connection between the sequence and the development of a specific disease was discovered. However, it can be said with a high probability that the three microRNAs differing from the rest are not false or a consequence of an experimental error. The actual reason of the obtained results is rather a biological or clinical difference between the molecules.

The most numerous group of microRNAs that cause nonspecific cardiovascular diseases was initially selected based on the heterogeneity of the formed cluster, due to its nonspecificity. Despite this, the results obtained indicate that all microRNAs in the group are located rather tightly to each other.

It is worth mentioning that during a detailed study of the

microRNA group in nonspecific cardiovascular diseases, it was not clear on what principle this group was created. Some microRNAs were the same as microRNAs in the group of cardiomegaly and coronary heart disease, which led to the idea that those molecules that could cause several cardiovascular diseases at once were collected in the group of nonspecific ones. However, in the course of further verification, this assumption was refuted.

One of the assumptions is that the formed group is in fact specific and has some common functions or clinical influence. In this case, a more detailed study of microRNAs located specifically in the group of nonspecific cardiovascular diseases is required.

It can also be assumed that the formation of dense groups can be carried out based not on function or clinical manifestations. Then there is some other connection between microRNAs in one group that does not affect the behavior of molecules in the human body.

Summing up, it is possible to determine further possible directions for the development of work:

1. Construction of a centroid for a cluster of microRNAs related to a specific disease. The centroid (the same vector of trimmer frequencies) will make it possible to form (generate) chains that are “average” for a given cluster (diagnosis). The created centroid would make it possible to predict possible microRNA sequences that can cause this disease in the future.
2. Using the method described in this work as a check of existing databases. Due to the large number of created databases, a way is needed to quickly check the classification, grouping of microRNAs.
3. Using the presented model to test new microRNA candidates for a given diagnosis (new candidates violate the homogeneity of the cluster, or vice versa, make it denser).

VI. Conclusion

The paper proposes a new method for assessing the clinical relationship of microRNA-diagnosis. It allows you to assess the quality of the relevant databases (find possible errors), as well as predict the presence of unknown microRNAs for selected diagnoses. The method is based on the comparison of microRNAs based on the frequency of occurrence of 3-mers.

The proposed model has been practically tested against the HMDD database. As a result of computational experiments for a group of heart diseases, all possible outcomes of the assessment of the classification of microRNA were confirmed - the diagnosis. Separate verified diagnoses (groups of microRNAs) formed a dense cluster, in some cases there was one single microRNA distant from the rest of the group, in other cases a subgroup of microRNAs was isolated (the existence of subclusters is shown).

For the clusters used in the experiment, an explanation of their inhomogeneity was proposed, as well as further possible directions for the development of this work in terms of prediction of microRNAs attributed to the selected diagnosis were determined.

References

[1] O'Brien, Jacob, et al. "Overview of microRNA biogenesis, mechanisms of actions, and circulation." Frontiers in endocrinology 9 (2018): 402.
[2] Chen, Hailin, et al. "Comparative analysis of similarity measurements in microRNAs with applications to microRNA-disease association predictions." BMC bioinformatics 21.1 (2020): 1-14.
[3] Çakmak, Hüseyin Altuğ, and Mehmet Demir. "MicroRNA and cardiovascular diseases." Balkan medical journal 37.2 (2020): 60.
[4] Yang, Yang, et al. "A clustering-based approach for efficient identification of microRNA combinatorial biomarkers." BMC genomics 18.2 (2017): 1-14.
[5] Liao, Jipei, et al. "MicroRNA- based biomarkers for diagnosis of non-small cell lung cancer (NSCLC)." Thoracic cancer 11.3 (2020): 762-768.
[6] Godlewski, Jakub, et al. "MicroRNA signatures and molecular subtypes of glioblastoma: the role of extracellular transfer." Stem cell reports 8.6 (2017): 1497-1505.
[7] Yoshimoto, Toyoki, et al. "Pulmonary carcinoids and low-grade gastrointestinal neuroendocrine tumors show common microRNA expression profiles, different from adenocarcinomas and small cell carcinomas." Neuroendocrinology 106.1 (2018): 47-57.
[8] Yang, Yi, and Xuting Wan. "Identification of MicroRNA Regulatory Modules by Clustering MicroRNA-Target Interactions." IEEE Access 8 (2020): 154133-154142.
[9] Askari Rad, Arezo, Jamal Fayazi, and Houshang Dehghanzadeh. "Clustering Some MicroRNAs Expressed in the Breast Tissue Using Shannon Information Theory and Comparing the Results With UPGMA, Neighbor-Joining, and Maximum-Likelihood Methods." Research in Molecular Medicine 8.4 (2020): 179-188.
[10] Fang, Wenwen, and David P. Bartel. "MicroRNA clustering assists processing of suboptimal microRNA hairpins through the action of the ERH protein." Molecular cell 78.2 (2020): 289-302.
[11] Kozomara, Ana, Maria Birgaoanu, and Sam Griffiths-Jones. "miRBase: from microRNA sequences to function." Nucleic acids research 47.D1 (2019): D155-D162.
[12] Maselli, Vincenza, Diego Di Bernardo, and Sandro Banfi. "CoGemiR: a comparative genomics microRNA database." BMC genomics 9.1 (2008): 1-9.
[13] Chen, Yuhao, and Xiaowei Wang. "miRDB: an online database for prediction of functional microRNA targets." Nucleic acids research 48.D1 (2020): D127-D131.
[14] Xie, Boya, et al. "miRCancer: a microRNA–cancer association database constructed by text mining on literature." Bioinformatics 29.5 (2013): 638-644.

[15] Jiang, Qinghua, et al. "miR2Disease: a manually curated database for microRNA deregulation in human disease." Nucleic acids research 37.suppl_1 (2009): D98-D104.
[16] Huang, Zhou, et al. "HMDD v3.0: a database for experimentally supported human microRNA–disease associations." Nucleic acids research 47.D1 (2019): D1013-D1017.
[17] Afshin Poorkhanalikoudehi, Karl-Heinz Zimmermann, "Cellular Automaton for Kidney Branching Morphogenesis." WSEAS Transactions on Biology and Biomedicine, 18 (2021): 170-182.