GPACDA – circRNA-disease association prediction with generating polynomials

Petr Ryšavý[1][0000−0002−6597−6616], Jiří Kléma[1][0000−0003−1753−9435], and Michaela Dostálová Merkerová[2][0000−0002−6345−9180]

[1] Department of Computer Science, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
[2] Department of Genomics, Institute of Hematology and Blood Transfusion, Prague, Czech Republic

Abstract. Circular RNA, a molecule with partially understood functions, has been implicated in various diseases. Therefore, there is a vast effort to predict associations between circular RNAs and diseases. In our recent study, we introduced circGPA, an algorithm that enables the annotation of circular RNAs with gene ontology terms through interactions with miRNAs and mRNAs. Recognizing the analytical similarity in predicting disease associations, we developed GPACDA, an extension of circGPA tailored for disease associations. The benefits of our methods include explainability, as the outputs are based on known interactions and associations, as well as the rigorous calculation of the p-value, which the circGPA algorithm can compute. We compared our method with two other tools, NCPCDA and DWNCPCDA, using a subset of the CDA-SOR dataset and showed that GPACDA overcomes its competitors in terms of true association ranks. Our method's code and predictions are publicly accessible.

1 Introduction

Circular ribonucleic acids (circRNAs) are single-stranded RNAs that, unlike linear RNAs, form a covalently closed continuous loop [43]. CircRNAs are widely expressed in eukaryotes in a tissue- and species-specific manner [7]. They already demonstrated their capacity to regulate gene expression and potentially link to diseases [50]. Owing to their remarkable stability, circRNAs can also serve as diagnostic biomarkers [50]. CircRNAs play a role in regulating gene expression by impacting transcription, mRNA turnover, and translation through interactions with RNA-binding proteins and microRNAs [40]. Annotation databases for circRNAs may encompass essential information regarding their tissue specificity, associations with diseases, and interactions with microRNAs [52]. Additionally, more advanced circRNA annotations can be derived from understanding their interactions with microRNAs and all other interactions involving these microRNAs [4]. Simultaneously, circRNA annotations may also originate from the known annotations of their host genes [28]. Nevertheless, the functions of most circRNAs remain unknown.

In this paper, we focus on the task of automated circRNA-disease association prediction. We define the problem as the classification of whether circRNA c of interest should be annotated by disease g. To do so, we exploit our recent tool circGPA [45], which uses statistic $s(c, g)$ calculated from the interaction graph between the circRNA, miRNAs, and mRNAs. Formally, statistic $s(c, g)$ is an outcome of the statistical test, whose null hypothesis is that the given (c, g) pair is unrelated. In other words, the null hypothesis is that circRNA c has no preference in interactions with miRNAs (or mediated interactions with mRNAs) associated with disease g. A circRNA c is annotated with a disease g if this null hypothesis is rejected. This principle assumes that miRNAs and mRNAs have already been partially explored and associated with most or at least some related diseases. As circGPA is based on generating polynomials, we name our novel method GPACDA (generating-polynomial annotator for circRNA-disease association prediction).

We have already demonstrated that circGPA is an efficient and exact method of circRNA annotation [45]. The main contribution of this paper lies in the application of circGPA to the new annotation task. Previously, we worked with ontology terms; now, we predict circRNA associations with diseases. This shift is by no means trivial, as it brings challenges in the construction of the interaction network and its annotation with known disease links. The disease vocabulary is much less established than the gene ontology (GO). Simultaneously, we have to consider the often missing disease associations; therefore, we have to look at them as incomplete and adjust the evaluation of the result accordingly. On the other hand, circRNA-disease annotation represents a more frequent task than previous circRNA-GO annotation. This allows us to benchmark our method better.

2 Relevant Work

Many tools focus on circRNA-disease association prediction. Most were produced in past years and are usually based on capturing similarities between circRNAs and diseases. In our experiments, we will compare against methods with publicly available circRNA-disease predictions. The first is NCPCDA [26], where the space of functional similarity of circRNAs is projected onto the circRNA-disease graph. Similarly, the semantic similarity of diseases is projected on the disease vertices. In the process, the similarities are combined with information about the known circRNA-disease associations. The network consistency projection is then used to build the final matrix of prediction scores.

The second reference algorithm is DWNCPDA [27], a tool based on DeepWalk and network consistency projection. The method calculates circRNA-circRNA similarities based on the known circRNA-disease associations using DeepWalk. Therefore, the method does not need any external biological input. A similar approach is used to calculate disease-disease similarities. Then, the network consistency projection is used to predict new associations similar to NCPCDA.

The schema employed in NCPCDA and DWNCPDA has been adopted by many other tools, each applying different methods to construct circRNA sim-

ilarities, disease similarities, and circRNA-disease associations. PWCDA [24] uses gene ontology terms that annotate circRNA-related genes, Jaccard index, and Gaussian interaction profiles. IBNPKATZ [60] combines bipartite network connections and KATZ measure. The CDASOR tool [32], which comes with a database that we will use for testing, employs the embedding of circRNA sequences into k-mer vectors and utilizes convolutional and recurrent neural networks to predict the associations. In contrast, the iGRLCDA tool [59] uses graph representation learning.

A popular path to tackle the circRNA-disease association prediction problem is to use a variety of machine learning methods. One of the first tools, ICFCDA [23], aimed at overcoming the sparsity of validated interactions by using recommender systems. AE-DNN [6] leverages autoencoders and deep neural networks. MLCDA [54] uses multi-source feature fusion. MDGF-MCEC [55] applies multi-view dual attention graph convolution network and cooperative ensemble learning. In a recent tool, THGNCDA [15], a triple heterogeneous graph network is utilized. Heterogeneous graph neural networks were also used in HGNCDA [29].

As the previous list suggests, the number of methods and techniques used is vast, and the review is far from complete. Other publications include [53,61,41]. Some of the papers [15,29,41] are from 2023, suggesting that the field is still undergoing active development. For readers with a deeper interest in the circRNA-disease association prediction field, we recommend review papers [22,5].

3 Previous Work – circGPA

We base our method on the recent circGPA algorithm [45]. The circGPA addresses a similar problem: the annotation of circRNAs with terms from the gene ontology. It is important to note a significant difference between the two problems: the frequency of known miRNA and mRNA associations with ontology terms is higher than disease associations.

The tool uses an interaction graph between the circRNA, miRNAs, and mRNAs. Let us denote the number of miRNAs (mRNAs, respectively) by $|\mu|$ ($|m|$). The circRNA of interest sponges miRNAs. This relation can be encoded in adjacency vector $\mathbf{a}^{\mu,c} \in [0,1]^{|\mu|}$ where each field equals 1 if the circRNA interacts with the corresponding miRNA, 0 otherwise. Similarly, matrix $\mathbf{A}^{m,\mu} \in [0,1]^{|m|\times|\mu|}$ contains 1 on the position corresponding to a pair of miRNA and mRNA whenever the miRNA silences the mRNA. The annotation of the miRNAs and mRNAs is stored in binary vectors $\mathbf{g}^{\mu} \in [0,1]^{|\mu|}$ and $\mathbf{g}^{m} \in [0,1]^{|m|}$. Denote $g = (\mathbf{g}^{m}, \mathbf{g}^{\mu})$.

The test statistic is then the number of paths of length one (two, respectively) from the circRNA to annotated miRNAs (mRNAs, respectively). Thus, the statistic is mathematically formulated as

$$s(c,g) = (\mathbf{a}^{\mu,c})^T \mathbf{g}^{\mu} + (\mathbf{A}^{m,\mu}\mathbf{a}^{\mu,c})^T \mathbf{g}^{m}. \tag{1}$$

The field in $\mathbf{a}^{\mu,c}$ ($\mathbf{A}^{m,\mu}\mathbf{a}^{\mu,c}$) corresponding to a miRNA (mRNA) will be called the *weight* of the miRNA (mRNA). Our tool circGPA [45] provides an efficient algorithm based on generating polynomials [11] to calculate the p-value

of $s(c, g)$. Define variables x and y. The power of x will count the weight, and the power of y will count the number of selected mRNAs. Focus now on a single mRNA with weight w. For this mRNA, we have two options: either it is not annotated with the term, resulting in polynomial x^0y^0, or it is annotated, resulting in $x^w y^1$. The sum of those is the contribution of the said mRNA to the statistic in Eq. (1). Also, we can notice that the coefficients by the polynomial terms show the number of ways to reach the value of the statistic with a term of a given size. For example, if we have two mRNAs with weights w_1 and w_2, the generating polynomial under the independence in the null hypothesis is

$$(1 + x^{w_1}y)(1 + x^{w_2}y) = 1 + (x^{w_1} + x^{w_2})y + x^{w_1+w_2}y^2. \tag{2}$$

From the polynomial, we see that there is one way to reach the statistic of 0 with an empty term, the statistic of w_1 or w_2 with a single element term, and the statistic of $w_1 + w_2$ with a two-element term.

Hence, we can define the generating polynomial for the statistic in Eq. (1) as

$$p(x, y, z) = \prod_{w \in \mathbf{A}^{m,\mu}\mathbf{a}^{\mu,c}} (1 + x^w y) \prod_{w \in \mathbf{a}^{\mu,c}} (1 + x^w z). \tag{3}$$

The z variable is used in the same meaning as y, but for the case of miRNAs.

The null distribution of statistic $s(c, g)$ is then found by the coefficients of polynomial $p(x, y, z)$, where the power of y is equal to $\|\mathbf{g}^m\|_1$ and the power of z is equal to $\|\mathbf{g}^\mu\|_1$. circGPA [45] then provides an efficient way to calculate the polynomial coefficients using dynamic programming. The algorithm exploits repeated integer weights by using the binomial expansion of repeated terms.

The technical details are beyond the scope of this paper. For now, it is important to mention the ,,guilt by association" principle [38], which posits that a circRNA associated with a disease tends to interact with miRNAs (and indirectly with mRNAs) that have already been associated with the disease. According to this principle, known associations from better-researched molecules are propagated to less-researched molecules, such as circRNAs.

4 GPACDA Methodology

GPACDA exploits the circGPA tool [45] designed originally to predict circRNA annotations with ontology terms. Now, we use the circGPA tool to predict the circRNA-disease associations. We reuse and update the circGPA interaction graph formally defined in Sec. 3. To obtain the annotation data, we exploit several databases with known mRNA/miRNA-disease associations and use them to annotate the interaction graph. The novelty lies in unifying data from several databases that provide disease associations for miRNAs and mRNAs and the proposal of a new formal method of circRNA-disease association prediction.

Interaction Graph The interaction graph we use in GPACDA is adopted from the circGPA tool [45]. The main difference is that we extended the database of

circRNAs (now, there are currently 4,555 circRNAs) and merged some of the miRNA IDs that represented the same sequence. As a result, the new, extended database that underlies the interaction graph contains 168,841 circRNA-miRNA interactions and 465,741 miRNA-mRNA associations.

The circRNA-miRNA interactions were downloaded from the CircInteractome [8] database that uses the TargetScan [25] interaction prediction algorithm. The miRNA-mRNA interactions were downloaded from the TarBase [20], miRecords [57], and miRTarBase [16] databases via the `multiMiR` package [44].

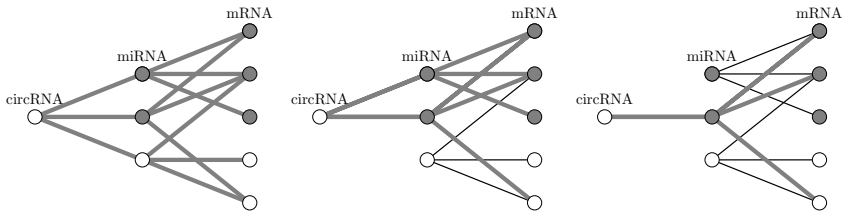
Known Disease Associations In this step, we collected the known miRNA- and mRNA-disease associations. The miRNA-disease associations were taken from the miR2disease database [19]. (available at `http://www.mir2disease.org/`) and the HMDD3 database [17]. Overall, a total of 38,499 miRNA-disease associations were available. The mRNA-disease associations were downloaded from the DisGeNET [42] database. The database is accessible through `disgenet2r` R package. This database provides 1,134,942 gene-disease associations and can be accessed programmatically by querying with selected diseases.

To construct the labeled interaction graph for the actual circRNA-disease pair, we picked a particular disease from a curated disease repository (for further details, see Sec. 5) and queried the above-described databases with it. As there is no widely adopted disease ontology, we had to rely on matching the names of the diseases in the databases. We used substring matching as a heuristic to retrieve the associated RNAs. For example, an mRNA known to be associated with *non-small cell lung cancer* will be included withing *lung cancer* disease.

Interaction Graph Examples To illustrate the behavior of GPACDA, we include Fig. 1 with three situations that can arise when GPACDA calculates the p-value of a circRNA-disease association. Neither the left nor the right image indicates an association between circRNA and disease, whereas the middle one does.

In Fig. 1a, the circRNA is densely connected to the rest of the graph. There are 12 paths from the circRNA to other molecules, and 8 of them terminate in a molecule associated with the disease. The circRNA-disease connection is, therefore, frequent but not statistically significant. The p-value is equal to 0.5, as in half the trials, the same or higher statistic would be reached if a random subset of mi/mRNAs of the same cardinality of 5 was selected.

The middle case in Fig. 1b indicates the association. Out of 8 paths, 7 terminate in an annotated molecule. Only two random sets of 5 mi/mRNAs reach $s(c, g) = 7$. In this case, the p-value is 0.067; on a larger graph, the p-value would be even smaller. The third case (Fig. 1c) occurs when the circRNA has a good connection to the disease. Nevertheless, the connection is supported by too few links. Three out of four paths terminate in a mi/mRNA associated with the disease. Still, there are many ways to reach the same number after randomly selecting mi/mRNAs – the circRNA is connected only to some of the mi/mRNAs with the known association, resulting in the p-value of 0.5.



(a) Dense connection. (b) Optimal connection. (c) Sparse connection.

Fig. 1: An illustration of the influence of the interaction graph structure on the resulting p-value. The gray nodes represent mi/mRNAs associated with the disease. The bold lines depict the paths from the circRNA to the disease nodes.

5 Experimental Evaluation

In this section, we define the ground-truth dataset, illustrate GPACDA outputs, and compare the predicted disease associations with reference algorithms. Since the problem is in the positive-unlabeled setting [2], the comparison is quite challenging. For evaluation, databases of known circRNA-disease associations are available, but there is no database of circRNAs that should not be associated with a disease. Therefore, only true-positive and false-negative predictions are known. The boundary between false-positive and true-negative is hard to evaluate, as a prediction not found in the database of known associations is not necessarily a false-positive but may be a valid circRNA-disease prediction missing in the database. This highlights the need for ongoing research efforts in circRNA-disease prediction algorithms.

Testing Disease List Overall, we work with a manually curated list of 60 diseases. The list was taken from the circ2disease [58]. It contains common diseases, including several types of cancer, atherosclerosis, multiple sclerosis, or rheumatoid arthritis. As there is no unified disease naming between the databases described in Sec. 4, we use the curated list just for their unification. We consider a disease equivalent to one of the 60 diseases if the substring relation holds (see Sec. 4). The circ2disease list also has significant overlap with our competitors. NCPCDA provides a list of 88 diseases, out of which 25 are in circ2disease. DWNCPCDA provides a list of 40 diseases, of which 18 are in circ2disease.

The Circ2disease is more consistent and more accessible to use than its counterparts. To illustrate the issue, the CDASOR [32] dataset contains terms with a broader range of granularity, often with possible overlaps. For example, there are diseases *cancer*, *lung cancer*, *lung carcinoma*, *small cell lung cancer*, *lung squamous cell carcinoma*, *non small cell lung carcinoma*, *non small cell lung cancer*, *lung adenocarcinoma* in CDASOR. The Circ2disease dataset, on the contrary, contains only half of the former list of diseases.

Projection of the Input Data on Testing Disease List After selecting this subset of diseases for testing, we end up with a cache of DisGeNET database [42] that

contains 5,895 mRNA-disease associations. After union with genes in the interaction graph, 3,921 associations remain. In the case of miRNAs, after selecting the 60 testing diseases and known miRNA IDs, only 90 associations are used. This rapid decrease in known miRNA associations is caused by poor overlap of miRNA IDs between HMDD3 and CircInteractome.

Projection of the Test Data on Testing Disease List For testing, we used the CDASOR dataset [32]. This dataset contains 3,221 ground-truth circRNA-disease annotations. Out of those, only 2,825 circRNAs are labeled with `hsa_circ_xxx` identifier used in CircInteractome. Next, for comparison, we select a subset of 501 associations used for testing. We required at least one miRNA (or mRNA, respectively) associated with the disease to interact (indirectly) with the circRNA of interest. Thus, the selected 501 associations represent the subset of the CDASOR dataset where the interaction graph for the circRNA and the disease associations for miRNAs and mRNAs, are known.

GPACDA Outputs In this section, we provide a brief illustration of the outputs generated by our algorithm. First, for a circRNA of interest, we present a sorted list of diseases. Each disease in the list is accompanied by a score statistic representing its association with the circRNA (as defined in Eq. (1)), the expected score under the assumption of independence, the number of mRNAs associated with the disease, and, most importantly, the corresponding p-value of the score statistic. Given that we are conducting multiple hypothesis testing, we also display the adjusted p-value using Bonferroni correction [9] and Holm FDR correction [3] for greater accuracy.

	$s(c,g)$	p-value	$\mathsf{E}\, s(c,g)$	$\|\mathbf{g}^m\|_1$	Bonferroni	FDR
acute myeloid leukemia	52	$1.6 \cdot 10^{-5}$	24	139	$8.7 \cdot 10^{-4}$	$8.7 \cdot 10^{-4}$
Alzheimer's disease	32	$2.0 \cdot 10^{-3}$	17	93	0.11	0.054
diabetic retinopathy	8	$4.5 \cdot 10^{-3}$	2	11	0.24	0.067
rheumatoid arthritis	50	$5.1 \cdot 10^{-3}$	32	175	0.27	0.067
glioma	36	$6.3 \cdot 10^{-3}$	21	118	0.34	0.067

Table 1: An example output of GPACDA on `hsa_circ_0000228`.

We illustrate the top predictions for circRNA `hsa_circ_0000228` in Tab. 1. The acute myeloid leukemia (AML) disease reaches a significantly low p-value. Based on the graph, we expect 24 paths to mRNAs and miRNAs associated with the disease on average, but `hsa_circ_0000228` has 52 connections to such mRNAs and miRNAs. There are 139 AML-associated mRNAs in total. Other diseases did not reach a significant p-value despite being close. For example, rheumatoid arthritis reached score 50, but it is a disease with more associations, resulting in a higher expected score and p-value. It is worth noting that `hsa_circ_0000228` is known to be connected with Myelodysplastic Syndromes [49], which in later stages can develop into AML. See Sec. 5 for more details.

Our method also provides output that helps with the explainability. With GPACDA, measuring the influence of individual miRNAs and mRNAs to final statistic $s(c, g)$ is possible. The derivation of the influence is in [45]. Such an output is in Fig. 2. The figure shows that the highest influence has `hsa-miR-194-5p`. This miRNA interacts with 13 mRNAs that are associated with AML. The genes that show the strongest influence are CCND2, CDK6, KMT2C, and RUNX1T1, all connected by three miRNAs to `hsa_circ_0000228`.

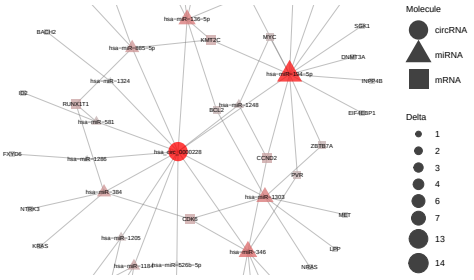


Fig. 2: A crop of graphical output of GPACDA showing the influence of individual miRNAs and mRNAs on the association prediction between `hsa_circ_0000228` and acute myeloid leukemia. The size of the nodes (delta) shows how much statistic $s(c, g)$ decreases if the mi/mRNA is removed from the interaction graph.

Reference Algorithms For reference, we compared our results with two benchmark algorithms, NCPCDA [26] and DWNCPCDA [27]. Those two tools published lists of ranked circRNA-disease association predictions, thus allowing easy comparison with other tools. Notably, the quantitative statistics available for GPACDA (see Section 5) are unavailable for NCPCDA and DWNCPCDA. As the results contain only the sorted lists of predictions, the direct comparison through true positive rate cannot be done in Section 5.

True Positive Rate In the first experiment, we compare the true positive rate on the testing dataset (see Sec. 5). As the problem is in the positive-unlabeled setting [2], providing a reliable ROC curve is challenging as only true positive and false negative data are available. Nevertheless, we show the dependence of the number of true-positive samples from the dataset based on the p-value threshold. The p-value threshold is, in this case, based on false discovery rate multiple hypothesis testing adjustment - the Benjamini-Hochberg method [3].

The results are in Fig. 3. The figure shows that at the level of significance 0.05, 195 associations are properly identified. However, with a significance of 0.1, half of the dataset (247 associations) is identified. On the contrary, the method did not provide adjusted p-value lower than one for 85 associations.

Rank of Ground Truth Predictions To compare the methods, we modify the outputs of GPACDA to provide a ranked list of circRNAs associated with a

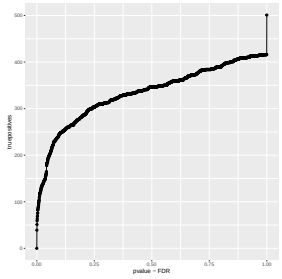
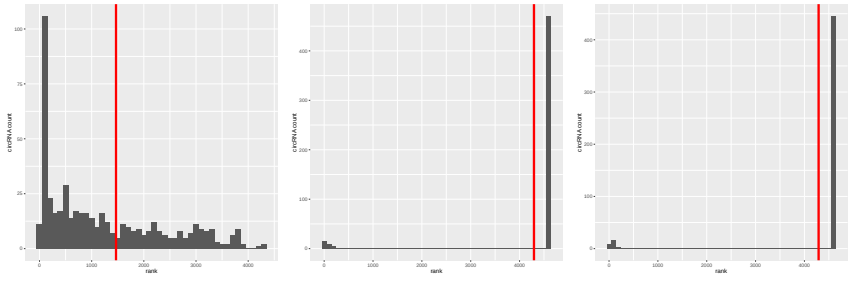


Fig. 3: Dependence of the number of ground-truth circRNA-disease associations labeled by GPACDA as positive based on the adjusted p-value threshold.

disease. For each test circRNA-disease pair (see Sec. 5), we calculate its rank. In the case of GPACDA, the rank of a circRNA-disease association is the number of circRNAs that have a lower p-value of the association with the same disease. In the case of NCPCDA and DWNCPCDA, we used the ranked list of associations provided as supplementary materials to the papers. Whenever an association from the test dataset was not present in the predictions, the rank was circRNAs.



(a) GPACDA (b) NCPCDA (c) DWNCPCDA

Fig. 4: The histograms of the ranks of ground-truth associations. For each association, the rank counts stronger predictions that share the same disease, i.e., a smaller rank means better predictions. The red line shows the average rank. For NCPCDA and DWNCPCDA, we include only diseases in the test dataset for which predictions are available. Hence, the histograms sum to less than 501.

The results are in Fig. 4. For NCPCDA and DWNCPCDA, only a few ground-truth circRNA-disease associations appeared in the predictions. In the case of GPACDA, 129 ($\sim$ 25 %) of the ground-truth dataset, are among the top 200 predictions. Since the problem is in the positive-unlabeled setting, ranking in the order of tens does not necessarily imply tens of false negatives above. Fig. 3 shows that half of the predictions reach a significant p-value, and thus, the rank is expected to be high.

Venn Diagram To illustrate the overlap between the methods, we include a Venn diagram of GPACDA, NCPCDA, and DWNCPCDA predictions. We set the FDR threshold to be equal to 0.05 in GPACDA. In the case of NCPCDA and DWNCPCDA, we include all predictions in the ranked lists. The Venn diagram is in Fig. 5. Results show a small overlap between methods, attributed partly to inconsistency in disease names and a lack of unified notation for circRNA IDs.

Fig. 5: Venn diagram of the predictions. The left plot shows all predictions by NCPCDA, DWNCPCDA, and GPACDA (restricted to the 60 diseases test list). The right plot shows the results on the diseases common for all three methods.

Case Study - the Myelodysplastic Syndromes (MDS) In this case study, we selected MDS, a group of cancerous diseases when blood cells in the blood marrow do not mature properly. In late stages, MDS can develop into acute myeloid leukemia (AML). We searched for circRNAs that can be related to MDS, presented the most reliable annotations, and verified the likelihood of our annotation against the literature. The method was initialized with 68 associations between MDS and genes. These are all the mRNA-MDS annotations in `disgenet2R`. There were also 76 miRNA-MDS associations in HMDD3. GPACDA predicted more than 1,000 circRNA-MDS pairs with FDR adjusted p-value smaller than 0.05. 20 top predictions with the lowest p-values are in Tab. 2.

The direct literature search for circRNA IDs in MDS context did not bring any hits; circRNA annotation is a relatively new task. Therefore, we included the host gene names and searched for evidence of their association with MDS. This search brought many overlaps between our predictions and the literature. For example, the second top prediction, circRNA `hsa_circ_0079009`, originates from the beta-tubulin gene (TUBB), a structural component of microtubules. It has been previously shown that TUBB gene expression was significantly higher in MDS patients who transformed to leukemia, and this gene may play a role in the leukemic transformation by affecting the proliferation of malignant clones [34,33].

Several circRNAs from the solute carrier (SLC) family group follow in the list. SLCs are membrane-bound proteins which play essential roles in a multitude of physiological and pharmacological processes. Perturbation of SLC transporter function underlies numerous human diseases and common genetic polymorphisms in SLC genes have been associated with inter-individual differences in drug efficacy and toxicity. Previous evidence suggested that some of the genes

from SLC protein family might be connected with drug resistance in MDS [63]. Also, the expression level of SLC25A1, one of the proteins of the family, has been associated with poor prognosis in AML [30].

Further, `hsa_circ_0007494` is produced from ROCK2 gene coding for a serin/threonine kinase. Paper [31] connects protein ROCK2 with NFKB pathway that contributes to many hematopoietic cell diseases. Similarly, GABPβ1 (host gene for `hsa_circ_0003501`), a GA-binding transcription factor, is necessary for myeloid differentiation and has a connection to chronic myeloid leukemia [35]. Additional evidence describing the relations of the remaining top twenty hit circRNAs with MDS and AML are referenced in Tab. 2. Only for three genes, no mentions with a connection to MDS nor AML have been described so far.

hsa_circ	FDR	Gene	Reference	hsa_circ	FDR	Gene	Reference
0048019	0,0011	ATP9B		0009140	0,0108	SCFD1	
0079009	0,0068	TUBB	[34,33]	0031423	0,0108	SCFD1	
0030045	0,0068	SLC25A15	*[63], [30]*	0006636	0,0113	PUM1	[37]
0084727	0,0068	SLCO5A1	*[63], [30]*	0001865	0,0131	UBQLN1	[47]
0001809	0,0075	SLCO5A1	*[63], [30]*	0002359	0,0132	UHRF2	[51]
0007494	0,0099	ROCK2	[31]	0072437	0,0132	PARP8	
0003501	0,0099	GABPB1	[35]	0067808	0,0132	RSRC1	[13]
0003715	0,0108	UBQLN1	[47]	0081083	0,0132	COL1A2	*[1]*
0081028	0,0108	PEX1	[36]	0006396	0,0142	BRAP	*[14]*
0041252	0,0108	PITPNA	*[12]*	0081084	0,0142	COL1A2	*[1]*

Table 2: Case study - 20 highest ranking predictions on MDS. References in italics show indirect connections to related genes or mentions of the genes.

In Tab. 3, we collected all circRNA-MDS pairs with a significant FDR and examined the frequency of circRNA-hosting genes. ITPR2, a key regulator for calcium ion transmembrane transportation activity, plays a critical role in cell cycle and proliferation. It has been proposed as a biomarker for worse prognosis and poor outcome in AML [46,56]. THBS1 might serve as a prognostic factor of AML; low expression levels of THBS1 indicate shorter overall survival [62]. Interestingly, MDS patients with bone marrow fibrosis showed increased expression of THBS1 [18]. Further, ABCC1 is a member of ATP-binding cassette transporters known to mediate chemotherapy resistance in AML and MDS [48]. High ABCC1 expression was also associated with poor disease-free survival [10].

Gene	ITPR2	THBS1	COL1A1	MGAT5	ABCC1
Frequency	23	21	10	7	7
References	[46,56]	[62,18]	[21]		[48,10]

Table 3: Case study - the genes hosting the most MDS-related circRNAs.

6 Discussion

GPACDA offers several advantages over its competitors. Firstly, it utilizes an exact algorithm for calculating p-value, allowing users to perform independent statistical assessments. When deciding whether to test a circRNA-disease connection experimentally, explainability is as crucial as the p-value. GPACDA, relying on the statistic $s(c, g)$, can generate informative visualizations, such as in Fig. 2 and provide insights into which interactions contribute to the low p-value.

GPACDA combines many inputs, being its strengths and weaknesses at once. Many sources allow GPACDA to eliminate possible data biases. Meanwhile, the databases from which disease annotations are taken contain many inconsistencies, as there is no widely adopted database of diseases similar to the gene ontology. When matching diseases, GPACDA has to rely on disease names, which are often at different levels of granularity. A unified database of known associations or a wider application of the disease ontology [39] could help future research.

The comparison of GPACDA and other tools shows that other tools often provide lists of relatively few predictions without comparability among different diseases. On the contrary, GPACDA can provide the p-value for any circRNA-disease pair. The list of predictions can be, therefore, arbitrarily long. Comparison with other diseases for a circRNA shows whether the disease is high in the list with a low p-value, high with a high p-value (a sparse interaction graph case), or low with a high p-value (an unlikely association). The tool choice should be based on available data for the circRNA of interest - GPACDA requires interactions, when they are not known, sequence based approaches might be better.

7 Conclusion

The annotation of circRNAs is an important task, given their emerging significance in molecular biology and their diagnostic potential. The detection and quantification of specific circRNAs can aid in early disease detection, monitoring treatment response, and predicting disease progression. Here, we introduced GPACDA, the efficient method for circRNA-disease association prediction. We would like to point out its two bold advantages: 1) it is capable of fast bulk evaluation of a large number of circRNA-disease pairs, and 2) it provides interpretable outcomes (each circRNA-disease pair can be explained with its interaction neighborhood, the molecules that most contributed to a positive association can be identified; the p-value assesses the strength of the evidence).

Our method effectively re-identifies known circRNA-disease associations, indicating potential for discovering new ones. This potential has further been reinforced with the MDS case study. GPACDA mostly detected circRNAs hosted by genes clearly related to the syndrome. The comparison with benchmark tools indicated that true circRNA-disease associations score higher in the GPACDA result lists than in the benchmark lists. GPACDA may fail, especially when the interaction graph or the existing m/miRNA-disease annotations is sparse. The results show that this setting does not occur frequently. The size of interaction

and annotation databases will continue to grow, which will only help to increase the precision and recall of the association prediction. Also, with growing number of validated interactions, GPACDA will be able to replace TargetScan predictions with validated circRNA-miRNA interactions.

GPACDA is available at github.com/petrrysavy/GPACDA with scripts to generate figures and input data downloaded from publicly available databases. Next, it might be worth to enrich the reasoning with gene expression data. Such integrative analysis would be beneficial in eliminating false positive associations by assigning low weights to interactions not supported by correlations in the expression data, thus, making expression-relevant associations to stand out.

Acknowledgment. This work was supported by the Ministry of Health of the Czech Republic - Czech Health Research Council, grant AZV NU20-03-00412.

References

1. Beau, M.M., et al.: Cytogenetic and molecular delineation of a region of chromosome 7 commonly deleted in malignant myeloid diseases. Blood **88**(6) (1996)
2. Bekker, J., et al.: Learning from positive and unlabeled data: a survey. ML (2020)
3. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. JRSS. Series B **57**(1) (1995)
4. Cardenas, J., et al.: Cerina: systematic circRNA functional annotation based on integrative analysis of ceRNA interactions. Scientific Reports **10**(1) (Dec 2020)
5. Chen, Y., et al.: Deep learning models for disease-associated circRNA prediction: a review. Briefings in Bioinformatics **23**(6) (09 2022)
6. Deepthi, K., Jereesh, A.: An ensemble approach for circrna-disease association prediction based on autoencoder and deep neural network. Gene **762** (2020)
7. Dong, R., et al.: CIRCpedia v2: An updated database for comprehensive circular RNA annotation and expression comparison. GPB **16**(4) (2018)
8. Dudekula, D.B., et al.: CircInteractome: A web tool for exploring circular RNAs and their interacting proteins and microRNAs. RNA Biology **13**(1) (2016)
9. Dunn, O.J.: Multiple comparisons among means. JASA **56**(293) (1961)
10. Ebner, J., et al.: Abcc1 and glutathione metabolism limit the efficacy of bcl-2 inhibitors in acute myeloid leukemia. Nature Communications **14**(1) (2023)
11. Feller, W.: Introduction to probability theory and its applications (1966)
12. Franzini, A., et al.: Molecular alterations in chronic myelomonocytic leukemia monocytes: Transcriptional and methylation profiling. Blood **132** (2018)
13. Gorombei, P., et al.: BCL-2 inhibitor ABT-737 effectively targets leukemia-initiating cells with differential regulation of relevant genes leading to extended survival in a NRAS/BCL-2 mouse model of high risk-MDS. IJMS **22**(19) (2021)
14. Guinn, B., et al.: Humoral detection of leukaemia-associated antigens in presentation acute myeloid leukaemia. BBRC **335**(4) (2005)
15. Guo, Y., Yi, M.: THGNCDA: circRNA–disease association prediction based on triple heterogeneous graph network. Briefings in Functional Genomics (09 2023)
16. Hsu, S.D., et al.: miRTarBase: a database curates experimentally validated microRNA–target interactions. Nucleic Acids Research **39**(suppl_1) (11 2010)
17. Huang, Z., et al.: HMDD v3.0: a database for experimentally supported human microRNA–disease associations. Nucleic Acids Research **47**(D1) (10 2018)

18. Hussein, K., et al.: Profile of fibrosis-related gene transcripts and megakaryocytic changes in the bone marrow of myelodysplastic syndromes with fibrosis. Annals of Hematology **97** (2018)
19. Jiang, Q., et al.: miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Research **37**(suppl_1) (10 2008)
20. Karagkouni, D., et al.: DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. NAR **46**(D1) (11 2017)
21. Kumar, B., et al.: Acute myeloid leukemia transforms the bone marrow niche into a leukemia-permissive microenvironment through exosome secretion. Leukemia **32**(3) (2018)
22. Lan, W., et al.: Benchmarking of computational methods for predicting circRNA-disease associations. Briefings in Bioinformatics **24**(1) (01 2023)
23. Lei, X., et al.: Predicting circRNA–disease associations based on improved collaboration filtering recommendation system with multiple data. Fr. in Genet. (2019)
24. Lei, X., et al.: PWCDA: Path weighted method for predicting circRNA-disease associations. International Journal of Molecular Sciences **19**(11) (Oct 2018)
25. Lewis, B.P., et al.: Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are MicroRNA targets. Cell **120**(1) (Jan 2005)
26. Li, G., et al.: NCPCDA: network consistency projection for circRNA–disease association prediction. RSC advances **9**(57) (2019)
27. Li, G., et al.: Potential circRNA-disease association prediction using DeepWalk and network consistency projection. Journal of Biomedical Informatics **112** (2020)
28. Li, S., et al.: Expression profile and bioinformatics analysis of circular RNAs in acute ischemic stroke in a South Chinese Han population. Sci. Reports **10**(1) (2020)
29. Liang, S., et al.: Hmcda: a novel method based on the heterogeneous graph neural network and metapath for circrna-disease associations prediction. BMC Bioinformatics **24**(1) (Sep 2023)
30. Liu, F., et al.: Slc25a1-associated prognostic signature predicts poor survival in acute myeloid leukemia patients. Frontiers in Genetics **13** (2023)
31. Liu, L., et al.: Mutated genes and driver pathways involved in myelodysplastic syndromes—a transcriptome sequencing based approach. Mol. BioSyst. **11** (2015)
32. Lu, C., et al.: Improving circRNA–disease association prediction by sequence and ontology representations with convolutional and recurrent neural networks. Bioinformatics **36**(24) (12 2020)
33. Ma, Y., et al.: The expression of beta-tubulin gene in myelodysplastic syndrome evoluting to leukemia. Zhonghua nei ke za zhi **55**(5) (May 2016)
34. Ma, Y., et al.: Prospective nested case–control study of feature genes related to leukemic evolution of myelodysplastic syndrome. Mol. Bio. Reports **40**(1) (2013)
35. Manukjan, G., et al.: Gabp is necessary for stem/progenitor cell maintenance and myeloid differentiation in human hematopoiesis and chronic myeloid leukemia. Stem Cell Research **16**(3) (2016)
36. Mo, G., et al.: Diagnostic approach to the evaluation of myeloid malignancies following car t-cell therapy in b-cell acute lymphoblastic leukemia. JITC **8**(2) (2020)
37. Naudin, C., et al.: PUMILIO/FOXP1 signaling drives expansion of hematopoietic stem/progenitor and leukemia cells. Blood **129**(18) (05 2017)
38. Oliver, S.: Guilt-by-association goes global. Nature **403**(6770) (2000)
39. Osborne, J.D., et al.: Annotating the human genome with disease ontology. BMC Genomics **10**(1) (Jul 2009)
40. Panda, A.C.: Circular RNAs: Biogenesis and Functions, chap. Circular RNAs Act as miRNA Sponges. Springer Singapore, Singapore (2018)

41. Peng, L., et al.: Predicting circrna-disease associations via feature convolution learning with heterogeneous graph attention network. IEEE JBHI **27**(6) (2023)
42. Piñero, J., et al.: DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database **2015** (04 2015)
43. Qu, S., et al.: Circular rna: a new star of noncoding rnas. Cancer letters (2015)
44. Ru, Y., et al.: The multiMiR R package and database: integration of microRNA–target interactions along with their disease and drug associations. Nucleic Acids Research **42**(17) (2014)
45. Ryšavý, P., Kléma, J., Merkerová, M.D.: circgpa: circrna functional annotation based on probability-generating functions. BMC Bioinformatics **23**(1) (Sep 2022)
46. Shi, J.l., et al.: High expression of inositol 1, 4, 5-trisphosphate receptor, type 2 (itpr2) as a novel biomarker for worse prognosis in cytogenetically normal acute myeloid leukemia. Oncotarget **6**(7) (2015)
47. Sweetser, D.A., et al.: Delineation of the minimal commonly deleted segment and identification of candidate tumor-suppressor genes in del(9q) acute myeloid leukemia. Genes, Chromosomes and Cancer **44**(3) (2005)
48. Szakács, G., et al.: Targeting multidrug resistance in cancer. NRDD **5**(3) (2006)
49. Trsova, I., et al.: Expression of circular rnas in myelodysplastic neoplasms and their association with mutations in the splicing factor gene sf3b1. Molecular Oncology
50. Verduci, L., et al.: CircRNAs: role in human diseases and potential use as biomarkers. Cell Death & Disease **12**(5) (May 2021)
51. Visconte, V., et al.: Splicing factor 3b subunit 1 (sf3b1) heterozygous mice manifest a hematologic phenotype similar to low risk myelodysplastic syndromes with ring sideroblasts. Blood **122**(21) (2013)
52. Vromman, M., et al.: Closing the circle: current state and perspectives of circular RNA databases. Briefings in Bioinformatics **22**(1) (01 2020)
53. Wang, L., et al.: MGRCDA: Metagraph recommendation method for predicting circRNA-disease association. IEEE Transactions on Cybernetics (2021)
54. Wang, L., et al.: A machine learning framework based on multi-source feature fusion for circRNA-disease association prediction. Briefings in Bioinf. **23**(5) (2022)
55. Wu, Q., et al.: MDGF-MCEC: a multi-view dual attention embedding model with cooperative ensemble learning for CircRNA-disease association prediction. Briefings in Bioinformatics **23**(5) (07 2022)
56. Wu, W., et al.: Characterization of bone marrow heterogeneity in nk-aml (m4/m5) based on single-cell rna sequencing. Exp. Hematology & Oncology **12**(1) (2023)
57. Xiao, F., et al.: miRecords: an integrated resource for microRNA–target interactions. Nucleic Acids Research **37**(suppl_1) (11 2008)
58. Yao, D., et al.: Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. Scientific Reports **8**(1) (Jul 2018)
59. Zhang, H.Y., et al.: iGRLCDA: identifying circRNA–disease association based on graph representation learning. Briefings in Bioinformatics **23**(3) (03 2022)
60. Zhao, Q., et al.: Integrating bipartite network projection and KATZ measure to identify novel circRNA-disease associations. IEEE TNB **18**(4) (2019)
61. Zheng, K., et al.: iCDA-CGR: Identification of circRNA-disease associations based on chaos game representation. PLOS Computational Biology **16**(5) (05 2020)
62. Zhu, L., et al.: Thbs1 is a novel serum prognostic factors of acute myeloid leukemia. Frontiers in Oncology **9** (2020)
63. Šimoničová, K., et al.: Different mechanisms of drug resistance to hypomethylating agents in the treatment of myelodysplastic syndromes and acute myeloid leukemia. Drug Resistance Updates **61** (2022)