Correlation and transcriptomic analysis revealing potential microRNA-gene interactions associated with breast cancer formation

Agustriawan David, Parikesit Arli Aditya*, Nurdiansyah Rizky, Ivan Jeremias and Ramanto Kevin Nathanael
Department of Bioinformatics, School of Life Sciences, Indonesia International Institute for Life Sciences, Jakarta, INDONESIA
*arli.parikesit@i3l.ac.id

Abstract
The results showed that the strongest negative correlation was peaked at -0.595 (p-value < 0.001); the miRNA and gene interaction was found between miR-1307 and CBX7. Then, transcriptomic validation result also supported the correlation.

Furthermore, the result of miR-1307 – CBX7 pair is also identified in the previous study through literature review. It suggests that the prediction accuracy is high. This study suggests that miR-1307 plays an oncogenic role and CBX7 potentially function as a tumor suppressor gene in breast cancer development. Taken together, correlation, transcriptomic and literature review study can be a rigorous combination pipeline to predict microRNA – gene interaction.

Keywords: Correlation, transcriptomic, breast cancer, molecular docking, aberrant miRNA, aberrant gene.

Introduction

Breast cancer is responsible for around 627, 000 deaths worldwide[17]. According to WHO report in 2018, breast cancer ranks fifth leading cause of death because the prognosis is relatively favorable. The incidence rate of female breast cancer is far exceeded in both developed and developing countries. In Indonesia, breast cancer case is rank one compared to other types of cancer[16]. Previous studies have shown that aberrant miRNAs expression levels involve in the tumorigenic process including breast cancer[17]. miRNAs are non-coding RNA that regulate important biological processes.

Furthermore, miRNAs have the ability to control gene expression by targeting mRNAs and start either translation repression or RNA degradation[1,8]. This suggests that miRNAs play a role as a novel class of oncogene or tumor suppressor.

In 2005, research that has been conducted by Iorio et al, compared the expression of miRNAs in breast cancer and normal cancer to identify miRNAs whose expression is significantly deregulated in cancer versus normal breast cancer. The result of 15 miRNAs expressions is able to correctly predict the nature of the sample analyzed which showed that aberrant expression of miRNA is involved in breast cancer. Another research conducted by Lee and Jiang[10] used a Bayesian Network (BN) to discover miRNA and gene possible interaction in breast cancer pathogenesis. The result discovered possible interaction of miRNA such as hsa-miR-21, hsa-miR-10b, hsa-miR-448 and hsa-miR-96 with oncogenes such as CCND2, ESR1, MET, NOTCH1, TGFBR2 and TGFB1 that involve in promoting tumor metastasis, invasion and cell proliferation.

In recent years, the records of interaction between miRNAs and gene interactions from wet lab analysis were stored in web-based miRNA related database such as miRTarBase[5]. These databases contain experimentally validated miRNA and gene interactions with detailed meta-data, experimental methodologies and conditions. Therefore, miRTarBase and other miRNA related database can be used to validate the miRNA-gene interaction. However, the last update of this website is on September 15, 2017, as we know, the high throughput sequencing technology such as next-generation sequencing (NGS) produces a massive amount of dataset each year. Therefore, the miRNA target interactions information need to be updated. Moreover, miRNA-target interactions list in miRTarBase is not based on a specific population.

Another method to validate the possible interaction between miRNA and gene is by using computer-aided discovery method such as molecular docking simulation. This method is used to investigate the binding and changes in the dynamic behavior of functional region miRNAs and gene[9]. Molecular docking is not only predicting the binding affinity of a ligand and receptor but also demonstrate the binding mode in the particular active site. Junaid et al[9] have performed docking analysis to DIM and miR-21that play a role in various cancer and other diseases. In the result, they showed that the stability of both DIM and miR-21 is negatively correlated to each other in binding condition.

Even though miRTarBase and transcriptomic analysis such as molecular docking and molecular dynamics can be used to discover the miRNA and gene interactions, finding the possible interaction in breast cancer is not an easy task. Most of the researcher used wet lab analysis in finding the possible interaction. However, this method is time-consuming, expensive and not efficient enough. Meanwhile, computational analysis approach has developed a different algorithm sometimes demanding a high requirement of a graphics processing unit (GPU) and big memory space[2].

Therefore, this study integrated correlation and transcriptomic analysis to overcome the problem of time and space complexity. This proposed method avoids all the

computation conducted in the transcriptomic analysis which has a problem in time and space complexity. The result of the correlation analysis will only be used for further transcriptomic analysis.

The increasing number of cancer and normal dataset both for miRNA and gene expression is recorded in TCGA website. More samples will generate a more valid prediction of a silico analysis. However, the number of dataset in a specific race and cancer stage is still limited, particularly for normal samples. Therefore, the main goal of this study is to propose a method to predict the interaction of aberrant miRNA with their target gene in breast cancer by utilizing the TCGA datasets. Spearman correlation analysis follows by molecular docking prediction as performed in this study. As a result, this study can provide a list of potential miRNA-gene pairs related to breast cancer formation with high accuracy. In a future study, a specific race and cancer stage will be included.

Material and Methods

Dataset: The RNA-seq gene expression datasets both normal and cancer patients were downloaded from TCGA on March 1st, 2019 and can be accessed in this following link https://portal.gdc.cancer.gov/. Breast invasive carcinoma (BRCA) cancer type was selected as a tumor model for the present work. The normalized gene expression unit used fragments per kilobase million – upper quartile (FPKM-UQ) and read per million (RPM) for gene and miRNA respectively. The expression units provide a digital measure of the abundance of transcripts.

Normalized expression units are necessary to remove technical biases in sequenced data such as depth of sequencing (more sequencing depth produces more read count for gene expressed at the same level) and gene length (differences in gene length generate unequal reads count for gene expressed at the same level; longer the gene more the read count). TCGA assembler package in R programming software was used for data retrieval. All the information, software and tutorial can be found in this link http://www.compgenome.org/TCGA-Assembler/.

Calculating differentially expressed genes and miRNAs: This study computed the gene expression average, standard deviation, the average + two standard deviations and the average – two standard deviations of healthy patients and also computed the average expression of cancer patients both in gene and miRNA dataset. In this study, certain genes or miRNAs are called aberrantly expressed if their expression in cancer patients is above or below the average expression of healthy patients +/- two times the standard deviation. All the calculation was computed using MySQL. As a result, differentially expressed gene (DEG) and differentially expressed miRNA (DEM) were determined.

Spearman Correlation computation: After this study determined the list of DEG and DEM, a correlation study was performed to identify the negative interaction of miRNA-gene pairs. There are two questions which can be answered by this correlation study. First, whether or not the down-regulation of DEG is induced by up-regulated DEM. Second, whether or not the up-regulation of DEG is affected by down-regulated DEM. First, this study conducted patient ID matching by TCGA2STAT package and OMICSBind function in R programming software for retrieving miRNA and gene expression dataset for the same set of cancer patients. For more information that function can be accessed through this link: http://www.liuzlab.org/TCGA2STAT/. Subsequently, this research developed a MATLAB code to compute Spearman correlation analysis, retrieved a rho and p-value from the correlation results. It yields a list of miRNA-gene pairs with statistically significant negative correlation.

Transcriptomic analysis: A list of miRNA – gene pairs was further analyzed using a transcriptomic procedure. First, this study obtained the sequences of miRNA and 3' UTR of miRNA-gene pairs from miRTarBase database which is an experimentally miRNA-gene interaction website that can be accessed in this link http://mirtarbase.mbc.nctu.edu.tw/php/index.php[5]. The current curation of miRTarBase is on September 15, 2017. It provides 8,510 number of articles, 23 number of species, 23,054 number of target genes, 4,076 number of miRNAs and 422,517 number of miRNA-target interactions.

The validation method was classified into two, the strong evidence included reporter assay, western blot and qPCR and less strong evidence included microarray, NGS, pSILAC. Second, the prediction of miRNA-target site was performed by RNAhybrid where it can predict the target site of miRNAs (https://bibiserv.cebitec.uni-bielefeld.de/rnahybrid)[15].

It is a tool to discover a minimum free energy hybridization of a long (target) and a short (query) RNA. The hybridization was conducted in a kind of domain model. The example is the short sequence hybridized to the best fitting parts of the long one. This study visualized the secondary and tertiary structures of the miRNA, gene and miRNA-mRNA duplex by RNAfold server that can be used through this link http://rna.tbi.univie.ac.at/cgibin/RNAWebSuite/RNAfold.cgi[12] and simRNAweb https://genesilico.pl/SimRNAweb/.[3,13] Lastly, molecular docking was performed to dock miRNA-mRNA duplex with Argonaute (AGO) protein (PDB ID:3F73, chain A) using PatchDock server (https://bioinfo3d.cs.tau.ac.il/PatchDock/).[4,7]

PatchDock is an algorithm to find the match between two molecules of any type of dataset included protein, DNA, peptides and drugs. Given two molecules, the algorithm works like assembling a jigsaw puzzle to match two pieces by picking one piece and searching for the complementary one. The complete procedure of the study is shown in figure 1.

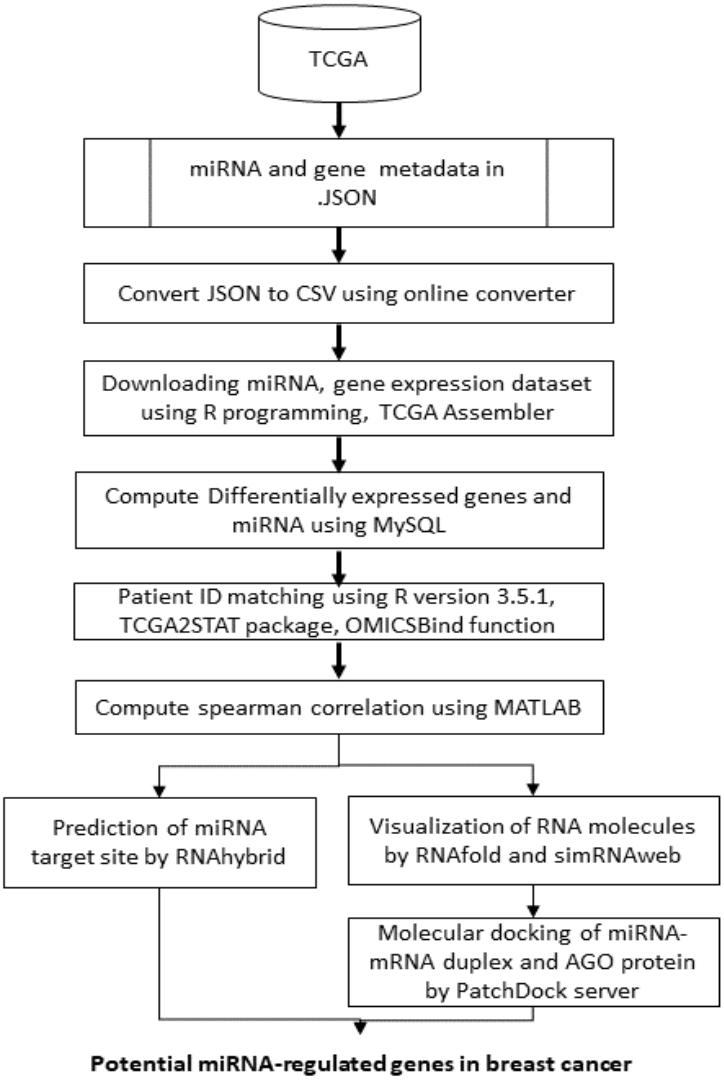

Figure 1: Flowchart of the present study.

Results and Discussion

A total of 1,095 cancer patients and 99 normal patient dataset were retrieved for gene expression. Moreover, a total of 765 cancer patients and 63 normal patient dataset were collected for miRNA expression. The high-throughput sequencing platform currently sequenced 20,531 and 1,880 genes and miRNAs respectively.

Given the dataset above, in order to discover a miRNA and gene biomarker which related to breast cancer formation, this study compared the mean expression of the cancer patients with the mean expression of the healthy +/- two standard deviations. From the normal distribution analysis, it indicates that the mean expression of the cancer datasets differs from the mean expression +/- two standard deviations from normal datasets.

This threshold can be used to determine whether the miRNAs and genes are upregulated or downregulated. This study found 122 and 3,735 downregulated genes and upregulated genes respectively. Moreover, 3 and 440 downregulated miRNA and upregulated miRNA were identified respectively. In total, there are 3,857 and 443 aberrant genes and miRNA respectively.

For those aberrant genes and miRNAs, a correlation computation was performed to infer a statistically negative significant correlation. It suggests that once the miRNAs expression is upregulated, it links with the downregulated gene expression and vice versa. We computed all the 440 upregulated miRNA with 122 down-regulated genes and furthermore, we also computed all 3 downregulated miRNAs with 3,735 upregulated genes. It yields miRNA-gene pair interactions. The correlation computation yields a Rho and P-value. For the threshold Rho < -0.2 with p-value < 0.05, 13,919 interactions were identified.

In order to filter the dataset, this study only considers a strong negative correlation, therefore a threshold with Rho < -0.5 and P-value <0.05 was selected. As a result, this study found 33 interactions as shown in table 1. All the miRNA in table 1 is an upregulated miRNA. Downregulated miRNA with their gene pair interactions is shown in table 2.

Table 1
Thirty-three most significant upregulated miRNA-down regulated gene pairs significantly negative correlation in breast cancer

miRNA	Gene	Rho	P-value
hsa-miR-1307	CBX7	-0.595	<0.0001
hsa-miR-210	CBX7	-0.590	<0.0001
hsa-miR-210	ITM2A	-0.589	<0.0001
hsa-miR-301a	CBX7	-0.567	<0.0001
hsa-miR-210	SCN4B	-0.567	<0.0001
hsa-miR-1307	SCNB4	-0.562	<0.0001
hsa-miR-1307	SPARCL1	-0.560	<0.0001
hsa-miR-130b	PCSK4	-0.558	<0.0001
hsa-miR-130b	MLPH	-0.555	<0.0001
hsa-miR-210	GIPC2	-0.542	<0.0001
hsa-miR-210	IL33	-0.539	<0.0001
hsa-miR-1301	SCB4B	-0.533	<0.0001
hsa-miR-190b	CHST3	-0.533	<0.0001
hsa-miR-210	SPARCL1	-0.533	<0.0001
hsa-miR-1301	CALCOCO1	-0.531	<0.0001
hsa-miR-130b	CALCOCO1	-0.529	<0.0001
hsa-miR-190b	A2ML1	-0.518	<0.0001
hsa-miR-210	GPRASP1	-0.516	<0.0001
hsa-miR-301b	CBX7	-0.515	<0.0001
hsa-miR-190b	PRNP	-0.515	<0.0001
hsa-miR-148b	CLEC11A	-0.512	<0.0001
hsa-miR-3677	CBX7	-0.512	<0.0001
hsa-miR-188	MLPH	-0.512	<0.0001
hsa-miR-135b	FOXA1	-0.511	<0.0001
hsa-miR-190b	KRT16	-0.511	<0.0001
hsa-miR-190b	GABBR2	-0.511	<0.0001
hsa-miR-1307	GPRASP1	-0.506	<0.0001
hsa-miR-135b	FSIP1	-0.506	<0.0001
hsa-miR-1307	CALCOCO1	-0.504	<0.0001
hsa-miR-130b	CBX7	-0.503	<0.0001
hsa-miR-1307	PDGFD	-0.502	<0.0001
hsa-miR-1307	JAM3	-0.501	<0.0001
hsa-miR-1301	CBX7	-0.501	<0.0001

According to table 1 and table 2, a miRNA can target more than one gene and vice versa. For docking analysis, this study will only consider the interactions in table 1 since we want to find the direct effect of miRNA toward gene expression. As a prediction model, the strongest negative correlation was selected for further analysis. miR-1307 and CBX7 have the strongest significantly negative correlation. This interaction could not be found in miRTarBase which experimentally validates the miRNA and gene interactions. Moreover, all the rest of the interactions was not recorded in miRTarBase. It means that it has not been validated yet in a wet lab experiment. miR-1307 and CBX7 pair were chosen as a model to assess the feasibility of the interaction in the transcriptomic analysis.

This study downloaded the sequence of miR-1307 and 3'UTR of CBX7 from miRTarBase, then miRNA-target site prediction was performed using RNAhybrid. Based on the result (figure 2), there was binding between CBX7 mRNA and miR-1307 at the first position of the UTR region denoted by 15 pairing nucleotides. Nucleotide bindings happen due to the hydrogen bonds between C-G (three bonds) and A-U (two bonds) of the interacting nucleotides.

Moreover, the minimum free energy (MFE) of the binding was -20.9 kcal/mol showing a favorable interaction. In addition, the sequences from figure 2 were shown in table 2.

The sequence information from table 3 was used to predict the secondary structure of all RNA molecules by RNAfold. Then, the dot-bracket notations as shown in table 4 were inputted into simRNAweb to predict the 3D structure of the RNAs. Furthermore, this study also downloaded the AGO protein sequence from the PDB database. The secondary structures of the RNAs were shown in figure 3 and the tertiary miRNA-mRNA duplex and AGO protein structure were shown in figure 4.

```
target 5'    U             A     3'
          ACG GCCGGUGUCG GC
          UGC UGGCUGCGGU CG
miRNA  3'               G  GCUCA 5'
```

Figure 2: Prediction of miR-1307 binding site on 3' UTR of CBX7 mRNA. The second line is the part of the first line (CBX7) and the third line is the part of the fourth line (miR-1307). The interacting nucleotides were placed near to each other and highlighted by yellow. U: uracil; A: adenine; G: guanine; C: cytosine

Lastly, using the 3d structure, the miRNA-mRNA duplex molecule was docked with AGO protein, a significant player in mRNA silencing process by using PatchDock Server. The first docking model was then retrieved with the statistical analysis and visualization in table 5 and figure 5 respectively. It shows that in PatchDock, the binding affinity is reflected from the score and ACE. As we took the conformation with the best score and negative value of ACE, the interaction must be favorable to happen.

Table 2
Top three of three downregulated miRNA-upregulated gene pairs significantly negative correlation in breast cancer

miRNA	Gene	Rho	P-value
hsa-miR-28	FAM174A	-0.406	<0.0001
hsa-miR-28	P4HTM	-0.404	<0.0001
hsa-miR-28	TMEM184A	-0.373	<0.0001
hsa-miR-10b	CDT1	-0.449	<0.0001
hsa-miR-10b	CDC45	-0.411	<0.0001
hsa-miR-10b	FAM64A	-0.407	<0.0001
hsa-let-7c	GINS2	-0.420	<0.0001
hsa-let-7c	PRIM1	-0.403	<0.0001
hsa-let-7c	UBE2T	-0.390	<0.0001

Table 3
Sequences of miR-1307 and its predicted target site on CBX7 mRNA

RNA	Sequences
Mature mir-1307	ACUCGGCGUGGCGUCGGUCGUG
Predicted miRNA target site	ACGUGCCGGUGUCGGCAUC
miRNA-mRNA duplex	GUGCUGGCUGCGGUGCGGCUCAA CGUGCCGGUGUCGGCAUC

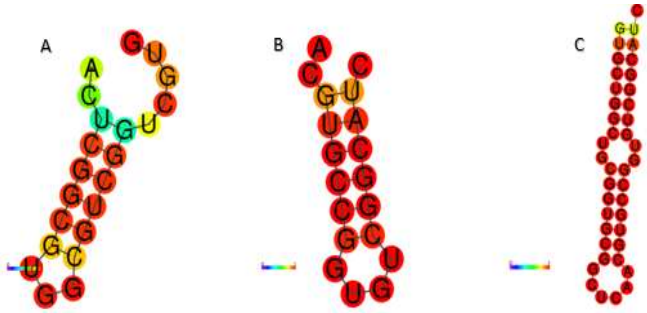

Figure 3: Secondary structure of RNA molecules. A: miR-1307; B: miRNA-target site of CBX7; C: miRNA-mRNA duplex. The colors denoted the conserved region with respect to the structure based on the base-pair probability parameter, from 0 (blue) to 1 (red)

Figure 4: The tertiary structure of RNA molecules. A: miRNA-mRNA duplex and B: AGO protein. The molecule was visualized by using PyMol version 2.3.

Table 4
Secondary structure of RNA molecules

RNA	Dot-bracket Notation	MFE (kcal/mol)
Mature mir-1307	...(((((...))))).....	-5.20
Predicted miRNA target site	..((((((....)))))).	-8.60
miRNA-mRNA duplex	(((((((((..(((((((.....))))))).)))))))).	-23.20

Abbreviation: MFE is denoted Minimum Free Energy

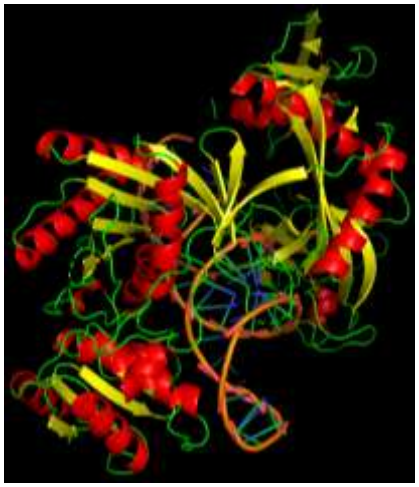

Figure 5: Docking result between miRNA-mRNA duplex with AGO protein. The molecule was visualized by using PyMol version 2.3

Table 5
Docking scores between miRNA-mRNA duplex with AGO protein

Molecules	Score	Area	ACE	Transformation
miR-1307 & CBX7-AGO	20576	3432.50	-703.10	0.60 0.65 -3.07 28.38 -3.73 39.12

Abbreviation: ACE is denoted as Atomic Contact Energy

Table 5 shows the best candidate complex between user-specified molecules. It suggests that miR-1307 interacts with AGO protein to target CBX7. The regulation of miRNA-gene pairs including the translation process and mRNA stability is affected by AGO protein [7]. AGO protein is the main actor of RNA-induced silencing complex (RISC) [4]. The domains in AGO protein anchor the 3' and 5' ends of miRNA and pointer it into SIRC. Furthermore, the AGO-centered RISC would bind to the 3' UTR of the target mRNA inhibiting the translation process[11]. This study finding also matches with previous study result; miR-1307 was found to be upregulated and plays an oncogenic role through targeting SET and MYND domain containing 4 (SMYD4) expression in breast cancer[6]. Polycomb protein family member CBX7 also played a critical role in cancer progression. It was found that reduced CBX7 protein levels were also observed in breast cancer carcinoma[14].

Taken together, this study identified that miR-1307 is an oncogenic miRNA which significantly contributes to the breast cancer tumor formation progression and inhibition and on the other hand, CBX7 could possibly is a tumor suppressor gene. Inhibition on miR-1307 can be a strategy to activate CBX7 which can cause the suppression of breast cancer initiation and progression.

Conclusion

Integrating correlation study and molecular docking analysis can be a novel method to accelerate the prediction of miRNA and gene interaction in cancer formation. The outcome of this study can suggest a further step that needs to be conducted to inhibit the potential oncogenic activity of microRNA or gene in breast cancer formation. For the future analysis, the silico study can be conducted by integrating another molecular biology such as DNA methylation which has been studied able to inhibit the expression of both coding and non-coding gene. For the final step, wet lab experiments can be performed, then the valid results can be designed for the treatment against cancer.

In a further study, this study will improve the methods to validate the aberrant miRNA and gene results. Subsequently, this study also will consider the effect of miRNA regulated gene in different races, cancer stage, gender etc. It is justifiable to be implemented since miRNA is one of the epigenetic cancer actors where its modification is caused by environmental factors.

Acknowledgement

The works of David Agustriawan is supported by the internal grant funding from the Research and Community Engagements Institute (LPPM) of Indonesia International Institute for life sciences (i3L). The authors would like to thank the LPPM and i3L for their supports.

References

1. Adams B.D., Kasinski A.L. and Slack F.J., Aberrant regulation and function of microRNAs in cancer, *Curr. Biol.*, **24(16)**, R762-R776 **(2014)**

2. Barbato C., Arisi I., Frizzo M.E., Brandi R., Da Sacco L. and Masotti A., Computational challenges in miRNA target predictions: to be or not to be a true target?, *J. Biomed. Biotechnol.*, DOI: 10.1155/2009/803069 **(2009)**

3. Boniecki M.J., Lach G., Dawson W.K., Tomala K., Lukasz P., Soltysinski T. and Bujnicki J.M., SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction, *Nucleic Acids Res.*, **44(7)**, e63-e63 **(2015)**

4. Carra S., Brunsting J.F., Lambert H., Landry J. and Kampinga H.H., HspB8 participates in protein quality control by a non-chaperone-like mechanism that requires eIF2α phosphorylation, *J. Biol. Chem.*, **284(9)**, 5523-5532 **(2009)**

5. Chou C.H., Shrestha S., Yang C.D., Chang N.W., Lin Y.L., Liao K.W. and Chiew M.Y., miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions, *Nucleic Acids Res*, **46(D1)**, D296-D302 **(2017)**

6. Han S., Zou H., Lee J.W., Han J., Kim H.C., Cheol J.J. and Kim H., miR-1307-3p stimulates breast Cancer development and progression by targeting SMYD4, *J. Cancer*, **10(2)**, 441 **(2019)**

7. Hutvagner G. and Simard M.J., Argonaute proteins: key players in RNA silencing, *Nat. Rev. Mol. Cell. Biol*, ***9*(1)**, 22 **(2008)**

8. Iorio M.V., Ferracin M., Liu C.G., Veronese A., Spizzo R., Sabbioni S. and Ménard S., MicroRNA gene expression deregulation in human breast cancer, *Cancer Res.*, **65(16)**, 7065-7070 **(2005)**

9. Junaid M., Dash R., Islam N., Chowdhury J., Alam M. J., Nath S.D. and Hosen S.Z., Molecular simulation studies of 3, 3′-Diindolylmethane as a Potent MicroRNA-21 Antagonist, *J. Pharm. Bioallied Sci.*, **9(4)**, 259 **(2017)**

10. Lee S. and Jiang X., Modeling miRNA-mRNA interactions that cause phenotypic abnormality in breast cancer patients, *PloS One*, **12(8)**, e0182666 **(2017)**

11. Li J., Kim T., Nutiu R., Ray D., Hughes T.R. and Zhang Z., Identifying mRNA sequence elements for target recognition by human Argonaute proteins, *Genome Res.*, **24(5)**, 775-785 **(2014)**

12. Lorenz R., Bernhart S.H., Zu Siederdissen C.H., Tafer H., Flamm C., Stadler P.F. and Hofacker I.L., ViennaRNA Package 2.0, *Algorithms Mol Biol.*, **6(1)**, 26 **(2011)**

13. Magnus M., Boniecki M.J., Dawson W. and Bujnicki J.M., SimRNAweb: a web server for RNA 3D structure modeling with optional restraints, *Nucleic Acids Res.*, **44(W1)**, W315-W319 **(2016)**

14. Pallante P., Forzati F., Federico A., Arra C. and Fusco A., Polycomb protein family member CBX7 plays a critical role in cancer progression, *Am. J. Cancer Res.*, ***5*(5)**, 1594 **(2015)**

15. Rehmsmeier M., Steffen P., Höchsmann M. and Giegerich R., Fast and effective prediction of microRNA/target duplexes, *RNA*, **10(10)**, 1507-1517 **(2004)**

16. The Global Cancer Observatory, Indonesia Summary Statistics, Retrived from http://gco.iarc.fr/today/data/factsheets/populations/360-indonesia-fact-sheets.pdf **(2019)**

17. WHO, Latest global cancer data: Cancer burdenrises to 18.1 million new cases and 9.6 milllion cancer deaths in 2018, Retrived from https://www.iarc.fr/wp-content/uploads/2018/09/pr263_E.pdf **(2018)**.

(Received 05th August 2020, accepted 10th October 2020)