Thyroid cancer tissues classification based on microRNA 3'-end modifications

Marta Danch, *Silesian University of Technology*

Abstract

Recent studies indicates that cellular cancerogenesis is connected with miRNA expression levels. In particular, different miRNAs can serve as classification features for distinguishing different cancer types. This paper provides classification attempt using miRNA isoforms with 3' end modification (see Table 1) as classification features. miRNA samples was obtained using next generation sequencing method. Data was preprocessed using authors algorithm developed in R. Support Vector Mashines and Partial Least Square methods were used to classify two types of miRNA samples: Follicular Adenoma and Follicular Thyroid Cancer (see Figure 1). It was also observed that only several miRNA modified isoforms were identified as the most differentiating for analyzed samples (see Table 2). Obtained results indicate that miRNA 3' end modifications can be used as cancer tissue classification features.

1. Introduction

Folicular Thyroid Cancer (FTC) constitute about 10% of all cancer diagnosis and is the second most common cancer type among people [12]. Unfortunately, distinguishing between FTC and its benign form Follicular Adenoma (FA) is extremely difficult. Based on histopathological studies only, classification efficiency varies from 11% to 69%. Large discrepancy is due to discrepancies in the data provided by various research centers. [12]

Based on genetic data, several models for tissue classification between FTC and FA were developed. For models classifying this type of tissue, in the literature one can find different performance indicators. These indicators are calculated only on the basis of a few selected miRNAs, and not all of the sequences present in the cells. Indicators in question can reach the level of about 80% for the accuracy [13] or 100% sensitivity and 86% specificity [4], depending on the data and feature selection methods. The best classification models differentiating between FTC and FA, for now, can reach the performance indicators up to 100% sensitivity and 94% specificity. However, these indicators were calculated basing on selected genes, not miRNAs [14]. These are highly satisfying results, but only from a statistical point of view. In fact, a certain number of patients still remains misdiagnosed.

miRNAs are short (15 - 27 base pairs), non-coding nucleotide sequences responsible for a number of mechanisms of controlling post-transcriptional processes in the cell. The study confirmed that miRNAs are involved in almost every cellular process. In mammals, they are responsible for about 50% of genes activity, for which they are specific inhibitors [5]. In case of humans more than 2 thousand miRNA sequences have been identified and described [10]. They are not only species specific, but also tissue specific. miRNA expression levels determines type of tissue origin and cell cycle stage. Based on miRNA presence and its expression levels one can identify acquired materials.

Recent three-year studies [5] has shown that miRNAs are not only involved in post-transcriptional gene processing, but they have also its own post-transcriptional processing path. Specific modifications were detected at the 3 'end of mature miRNAs, which consist of up to 3 additional bases [11], with the most frequent modification is adenine (A) [15, 11, 7]. Additional bases do not come from unmatures miRNA (hairpin) sequences. Exemplary modified sequences shown in Table 1.

In tested samples specific diversity of modification was detected, which rules out random base addition. [6]. Modification processes are not yet fully understood, but it is known that these processes are biologically regulated [1, 15].

Tab.1.

Exemplary isoforms table with occurrence number and their percentage share for miRNA sample (here: hsa-miR-486-5p). Modification bases detected in data are labeled with bold font.

	Isoform	Amount	%
miRNA reference	TCCTGTACTGAGCTGCCCCGAG		
detected in data miRNA sequences	TCCTGTACTGAGCTGCCCCGAG	1042	51,20%
	TCCTGTACTGAGCTGCCCCG	15	0,74%
	TCCTGTACTGAGCTGCCCCGA	217	10,66%
	TCCTGTACTGAGCTGCCCC	2	0,10%
hairpin reference	ATCCTGTACTGAGCTGCCCCGAGGCCCT		
detected in data hairpin sequences	TCCTGTACTGAGCTGCCCCGAGG	6	0,29%
	ATCCTGTACTGAGCTGCCCCGA	1	0,05%
	ATCCTGTACTGAGCTGCCCCGAG	1	0,05%
	TCCTGTACTGAGCTGCCCCGAG**A**	591	29,04%
	TCCTGTACTGAGCTGCCCCGAG**T**	124	6,09%
	TCCTGTACTGAGCTGCCCCGAG**AT**	9	0,44%
	TCCTGTACTGAGCTGCCCCGAG**GT**	2	0,10%
	TCCTGTACTGAGCTGCCCCGAG**GAG**	1	0,05%
	TCCTGTACTGAGCTGCCCCGAG**TA**	5	0,25%
	TCCTGTACTGAGCTGCCCCGAG**GA**	1	0,05%
	TCCTGTACTGAGCTGCCCCGAG**AA**	10	0,49%
	TCCTGTACTGAGCTGCCCCGAG**C**	1	0,05%
	TCCTGTACTGAGCTGCCCCGAG**TT**	1	0,05%
	TCCTGTACTGAGCTGCCCCGAG**AGA**	1	0,05%
	TCCTGTACTGAGCTGCCCCGAG**TAA**	1	0,05%
	TCCTGTACTGAGCTGCCCCGAG**AG**	1	0,05%

At the moment, there are few reports on the biological function of 3'end miRNAs modification. It is supposed, that modifications are stabilizing part of miRNAs structure, so it is not immediately degraded [5]. Since many researches confirm species and also tissue diversity of 3'end modifications, it is not possible to create a global modification pattern [1, 6, 7, 11, 15]. Thanks to this, there is a possibility that the expression levels of modified miRNAs or characteristic modifications could constitute set of features allowing tissue classification.

2. Materials and Methods

The study used 20 miRNA samples from the FTC and FA tissues obtained by deep sequencing method (often referred as next generation sequencing - NGS) [3]. This technology, from year to year more and more popular, using fluorescent additives to nucleic acids, allows detection of all available cell genetic material in a relatively short period of time. In cases when sequences under test repeat themselves, like RNA sequences, basing on data from NGS one can specify the levels of expression of these sequences, and thus the approximate concentrations of proteins present in the tested cells.

Sequences detected in samples have been processed using author's algorithm in which miRNA isoforms with 3'end modification were selected. Created algorithm is based on a bowtie2 and R scripts. Obtained expression levels for each isoform were normalized in order to be able to compare values for individual samples.

Expression levels of individual miRNA isoforms detected in each sample were used as features set for classification models. Classification models were created using supervised methods of analysis: Support Vector Machine (SVM) [2] and the Partial Least Squares (PLS) [9] method. For model validation boostrap632 method (combination of classic bootstrap and resubstitution method) was used and set of 10 differentiating features was selected from the data using Wilcoxon test. In order to determine the quality of the classification three most common indicators were used: sensitivity, specificity, and accuracy. The whole analysis was performed using libraries and R language scripts.

3. Results

As shown in the Figure 1 classification based on PLS obtained higher values of quality indicators than SVM method

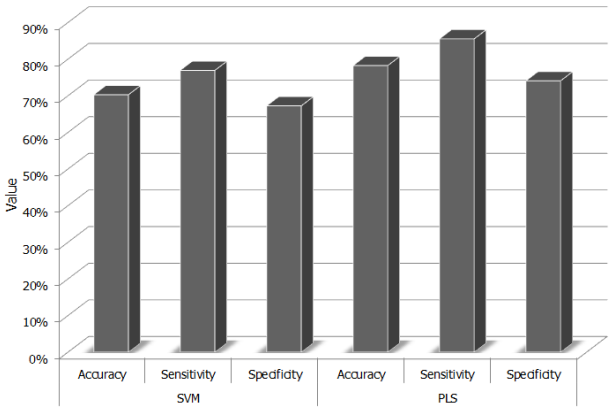

Fig.1. Obtained performance indicators for classification methods used in the paper.

In analyzed data 424 miRNA isoforms with 3' end modification were identified. For purpose of creating models about 50% of them were used.

To create classification models, set of five most often used isoforms were selected. Table 2 presents isoforms used for classification and percentage utilization in models for different validation methods.

In case of bootstrap validation method presented features are those with the largest percentage of use, but the values are much lower than in case of resubstitution method. Because the resubstitution method uses more features than bootstrap method to build classification model, such model is more complete, and therefore more reliable.

Polinucleotide modified isoforms were the most differentiating features in created

classification models, despite singlenucleotide modified isoforms have the highest expression levels in used samples.

Tab.2
Most commonly used isoforms in classification models. Abbreviations stands for validation methods: boot – bootstrap method, res – resubstitution method. Modification bases detected in data are labeled with bold font.

miRNA reference	Isoform	Percentage usage [%]	
		boot	res
hsa-mir-30e--5p	TGTAAACATCCTTGACTGGAAG**CTT**	64%	100%
hsa-mir-21--5p	TAGCTTATCAGACTGATGTTGA**CC**	59%	100%
hsa-mir-486--5p	TCCTGTACTGAGCTGCCCCGAG**AG**	51%	100%
hsa-mir-222--3p	AGCTACATCTGGCTACTGGGTCTC**A**	43%	100%
hsa-mir-28--3p	CACTAGATTGTGAGCTCCTGGA**AA**	37%	100%

4. Discussion

Based on the results obtained in the work one can conclude that modifications of 3' end miRNA work well as a feature classifying the different types of tissues. Modified miRNA isoform expression levels differentiates two classes of samples at a satisfactory level. This confirms thesis cited in the introduction [1, 6, 7, 11, 15] that modifications are tissue-specific.

Modified isoforms detected in data suggest that these modification sequences are not random. According to what was previously described [7, 11, 15], the most common modification is Adenine. Not all modified isoform expression levels have statistically significant differences, what confirms omitting a large part of them during creating classification models. It is highly possible that those sequences are characteristic for thyroid tissue or general malignancy indicators. Comparing obtained results with different set of data from the above mentioned thyroid tissue, could assist in confirm or overthrow this hypothesis.

In literature several differentiating sequences are described. Among the most frequently mentioned, there are hsa-mir-21, hsa-mir-192, hsa-mir-197, hsa-mir-222, hsa-mir-328, hsa-mir-346 [4, 8, 12, 13]. Two of above mentioned miRNA isoforms (hsa-mir-21, hsa-mir-222) were also identified in this paper, as the most differentiating for analyzed samples (see Table 2).

In this paper one of the simplest and most intuitive feature selection test(non-parametric Wilcoxon test) was used. There is a possibility to define better classification model using another more precised method. Another alternative method of feature selection could base on biological approach. Knowledge of thyroid cancer specific miRNA could be also good way to select differentiating FA and FTC isoforms.

5. Conclusions

It is not surprising, that classification models based only on modified isoforms cannot reach 100% accuracy with thyroid cancer tissues classification. Nonetheless, 70-80% efficiency of classification allows to locate the 3' end miRNA modification between other markers, such as individual genes and miRNAs. One classification model created basing on all these features i.e. genes, miRNAs, modified miRNAs, could improve already existing schemes and provide reliable classifications.

Classification problem mentioned in this paper was solved basing on twenty tissue samples. Supposing that satisfactory obtained results are not accidental, additional amount of data could clarify created models, and thus improve classification quality.

Significant growth of interest in the 3'end miRNA modification shows up in growing amount of publications, providing information that could complement this paper with new facts and probably extend data set required for further classification problem.

6. Acknowledgements

This work was supported by Silesian University of Technology.

7. References

1. Burroughs A. Maxwell et al.: *A comprehensive survey of 3' animal mirna modification events and a possible role for 3' adenylation in modulating mirna targeting effectiveness*, Genome Research, 2:1398-1410, 2010

2. Dudoit Sandrine and Fridlyand Jane: *Classification in microarray experiments*, Bioconductor Manual, September 2002.

3. Illumina company website: *www.illumina.com*

4. Keutgen Xavier M et al.: *A panel of four micrornas accurately differentiates malignant from benign indeterminante thyroid lesions of fine needle aspiration*, Clinical Cancer Research, Published Online 20.02.2012, 2012

5. Krol Jacek et al.: *The widespread regulation of microrna biogenesis, function and decay*, Nature Reviews, 11, 2010

6. Lee Lik Wee et al.: *Complexity of the microrna repertoire revealed by nextgeneration sequencing*, RNA, 16:2170-2180, 2010

7. Li Sung-Chou et al.: *Microrna 3' end nucleotide modification patterns and arm selection preference in liver tissues*, In 23rd International Conference on Genome Informatics GIW 2012, 2012

8. Marini Francesca et al.: *Microrna role in thyroid cancer development*, Journal of Thyroid Research, 2011 ID: 407123, 2011
9. Mevik Bjorn-Helge and Wehrens Ron: *The pls package: Principal component and partial least squares regression*, Journal of Statistical Software, 18, 2007
10. Mirbase database website: *www.mirbase.org*
11. Newman Martin A. et al.: *Deep sequencing of microrna precursors reveals extensive 3' end modification*, RNA, 17, 2011
12. Rossing Maria et al.: *Classification of follicular cell-derived thyroid cancer by global rna profiling*, Journal of Molecular Endocrinology, 2013
13. Weber Frank et al.: *A limited set of human microrna is deregulated in follicular thyroid carcinoma*, The Journal of Clinical Endocrinology and Metabolism, 91(9):3584-3591, 2006
14. Weber Frank et al.: *Genetic classification of benign and malignant thyroid follicular neoplasia based on a three-gene combination*, The Journal of Clinical Endocrinology and Metabolism, 90(5):2512-2521, 2005
15. Wyman Stacia K. et al.: *Post-transcriptional generation of mirna variants by multiple nucleotidyl transferases contributes to mirna transcriptome compexity*. Genome Research, 21:1450-1461, 2011

Author:

MSc. Marta Danch
Silesian University
of Technology,
Faculty of Automatic
Control, Electronics and
Computer Science
ul. Akademicka 16
44-100 Gliwice
Tel: (+48) 32 23 72 309

email: *marta.danch@gmail.com*