Computers in Biology and Medicine 64 (2015) 261–267

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Analysis of shared miRNAs of different species using ensemble CCA and genetic distance

Nazife Cevik [a,*], C. Okan Sakar [b], Olcay Kursun [a]

[a] Department of Computer Engineering, Istanbul University, Istanbul 34320, Turkey
[b] Department of Computer Engineering, Bahcesehir University, Istanbul 34353, Turkey

ARTICLE INFO

Article history:
Received 6 April 2015
Accepted 24 June 2015

Keywords:
Canonical correlation analysis
Ensemble methods
miRNA sequence analysis
Multivariate statistics
Genetic distance

ABSTRACT

MicroRNA is a type of single stranded RNA molecule and has an important role for gene expression. Although there have been a number of computational methodologies in bioinformatics research for miRNA classification and target prediction tasks, analysis of shared miRNAs among different species has not yet been addressed. In this article, we analyzed miRNAs that have the same name and function but have different sequences and belong to different (but closely related) species which are constructed from the online miRBase database. We used sequence-driven features and performed the standard and the ensemble versions of Canonical Correlation Analysis (CCA). However, due to its sensitivity to noise and outliers, we extended it using an ensemble approach. Using linear combinations of dimer features, the proposed Ensemble CCA (ECCA) method has identified higher test-set-correlations than CCA. Moreover, our analysis reveals that the Redundancy Index of ECCA applied to a pair of species has correlation with their genetic distance.

1. Introduction

MicroRNA (miRNA) is a type of single stranded RNA molecule that is approximately 21–23 nucleotides long and responsible from the arrangement of gene expressions [1–6]. miRNAs are non-coding RNAs, i.e. they are coded by genes that are not translated into proteins but is transcribed by DNA [7]. Primer transcripts known as pri-miRNAs are processed and first transformed into pre-miRNA hairpin loop and then functional miRNA. Researchers have reported in humans that miRNAs regulate many fundamental cellular functions, therefore the abnormal levels of miRNA levels in cells is linked to the development of cancer [8,9]. The effect of miRNAs as oncogenes or tumor suppressors during tumor development stages has also been addressed [10,11]. Therefore, a working knowledge of miRNAs is very important for early detection and treatment of cancer and may be associated with several other diseases [12].

In the literature, most of the machine learning studies dealing with miRNAs addresses the identification of miRNA target predictions and mRNA target sequences by using classification techniques such as Naive Bayes or Support Vector Machines (SVMs). Malik et al. [13] used Naive Bayes classifier to predict miRNA target predictions. As input to Naive Bayes classifier, they proposed to use features extracted from both the seed and 'out seed' sequences and the duplex structures. A big amount of mature miRNA and a collection of several confirmed miRNA targets were used to generate the positive and negative classes. They concluded that the seed segment is the most important factor in target selection. Thus, because out-seed segment has an important role in the target interaction for miRNAs, both seed and out-seed regions of the miRNA sequence are included in their feature set. Chenghai [14] aimed to classify real pre-miRNA by feeding the combination of sequence and structure information of stem loops extracted from hairpin sequences as input to SVMs. In [15], a cancer classification method that aims to detect some abnormal miRNA expression patterns was proposed. They used a mutual information based method to evaluate the subsets of miRNAs. In another study [16], a classification method based on pre-miRNAs was proposed for the predicton of human miRNA gene. The main idea of this approach is to identify pre-miRNAs among hairpin structure of the human genome. The dataset contains ncRNA and pseudo hairpins as one class and human pre-miRNA sequences and non-redundant sequences folded into hairpin secondary structure as the other. Several sequence and folding features have been used to discriminate these classes by an SVM classifier. To improve the accuracy of these studies, ensemble learning, a recently popular technique that combines multiple models to finally obtain a robust model, has also been recently used in miRNA classification tasks [17]. As an application of Canonical Correlation Analysis (CCA) to miRNAs, the study in [18] looks for the phylogenetic

* Corresponding author.
E-mail address: ncevik@istanbul.edu.tr (N. Cevik).

http://dx.doi.org/10.1016/j.compbiomed.2015.06.023
0010-4825/

dependence between the investigations of biological species. Another approach has used CCA for both univariate and multivariate gene-based tests of association [19].

Although a large number of machine learning methods have been effectively used for analyzing miRNA sequence features for classification and target prediction tasks, the analysis of shared miRNAs among species has not yet been addressed in the literature. Moreover, instead of using the classical CCA, to avoid its shortcomings, we used our proposed ensemble CCA method [20]. Traditionally, sequence based features such as Minimum Free Energy (MFE), individual base frequencies, individual dimer frequencies, or percent GC content, %(G+C), can all express some correlation between common miRNAs of different species; however, in this paper we show that CCA and ECCA can better identify maximally correlated weighted/linear combinations of these sequence features.

The rest of this paper is organized as follows. In Section 2, we firstly describe the miRNAs dataset used for our analyses, and then give a brief review of classical CCA and our proposed Ensemble CCA method. We present the simulation results in Section 3. Finally, we conclude with a discussion of our results in Section 4.

2. Materials and methods

2.1. miRNA genomic sequence datasets

Shared miRNAs have the same name and function but have different sequences and belong to different (but closely related) species. miRNA precursors used in the analysis belong to different species (15 in total) were derived from mirBase which is a searchable database for all published miRNA sequences and occurrence [21–24]. Each entry of sequence database represents hairpin portion of miRNAs with information on the location and sequences of the pre-miRNA and mature miRNA. We first analyze shared miRNA precursors among the species (see Table 1). We chose some species pairs that have a descent number of shared miRNAs (>90). Along with the species that are siblings in the biological tree, we included closely related non-sibling species pairs as well.

The number of shared miRNAs between the analyzed species pairs demonstrates that species pairs with sibling relationship have more common miRNAs than the others (see Table 2).

For instance, when let-7a-1 miRNA precursor sequences of Equus caballus and Canis familiaris species are aligned (see Fig. 1), it is seen that shared miRNAs have similar sequences but not exactly the same. Hence, the nature and extent of relations between them need exploration.

Table 1
Number of miRNAs of species.

Species	Class	Order	# of miRNAs
Bos taurus	Mammalia	Artiodactyla	766
Cricetulus griseus	Mammalia	Rodentia	200
Homo sapiens	Mammalia	Primates	1600
Sus scrofa	Mammalia	Artiodactyla	271
Equus caballus	Mammalia	Perissodactyla	341
Anolis carolinensis	Reptilia	Squamata	282
Canis familiaris	Mammalia	Carnivora	323
Monodelphis domestica	Mammalia	Didelphimorphia	156
Taeniopygia guttata	Aves	Passeriformes	243
Drosophila grimshawi	Insecta	Diptera	82
Drosophila melanogaster	Insecta	Diptera	238
Branchiostoma floridae	Leptocardii	Amphioxiformes	156
Ovis aries	Mammalia	Artiodactyla	105
Petromyzon marinus	Petromyzontida	Petromyzontiformes	244
Tribolium castaneum	Insecta	Coleoptera	220

Table 2
Number of shared miRNAs between species pairs.

Species	# of Shared miRNAs
Bos taurus – Cricetulus griseus	143
Bos taurus – Homo sapiens	226
Bos taurus – Sus scrofa	172
Equus caballus – Anolis carolinensis	98
Equus caballus – Canis familiaris	222
Equus caballus – Homo sapiens	226
Equus caballus – Monodelphis domestica	95
Monodelphis domestica – Canis familiaris	102
Canis familiaris – Anolis carolinensis	114
Bos taurus – Anolis carolinensis	139
Bos Taurus – Taeniopygia guttata	102
Anolis carolinensis – Taeniopygia guttata	103
Anolis carolinensis – Homo sapiens	132
Anolis carolinensis – Sus scrofa	92
Taeniopygia guttata – Homo sapiens	97

2.2. Definitions

- Let P_i denotes Jukes–Cantor genetic distance between miRNA*i* of two species; Common miRNA Distance(CMD) is defined as

$$CMD = \frac{\sum_{i=1}^{n} P_i}{n}$$

- Let *N*1 is the number of miRNAs for the first species and *N*2 is the number of miRNAs for the second species among species pairs.
- Given a species pair, Let N denotes the number of shared miRNA $\{miRNA_1 \ldots. miRNA_N\}$ and P*i* denotes Jukes–Cantor genetic distance between miRNA*i* of the two species.
- Let A is the set of miRNAs for the first species and B is the set of miRNAs for the second species among species pair, then Jaccard Index(JI) is as follows:

$$JI = \frac{|A \cap B|}{|A \cup B|} = \frac{N}{N_1 + N_2 - N}$$

2.3. Algorithm for correlation between common miRNA distance (CMD) and redundancy index (RI)

The following algorithm is used to obtain correlation between Common miRNA Distance which we call CMD and Redundancy Index(RI).

Input: miRNA*i* sequences of set A and miRNA*i* sequences of set B, then, F*A* is the feature extracted from set A and F*B* is the feature extracted from set B.

Output: Correlation between CMD and RI

1. Set $i=1$
2. while $i \leq N$ do
3. sequences{1,1}=miRNA*i* of set A
4. sequences{2,1}=miRNA*i* of set B
5. *Pi*=seqpdist(sequences)
6. stop
7. Get CMD
8. Obtain CCA on F*A* and F*B*
9. Set $n=1$
10. while $n \leq$ dim
11. Obtain RI for canonical cross loading
12. stop