Potential MiRNAs Recognition Site Identification in 3' UTR Regions by DSP Methods

Norbert Maggi, *Member, IEEE*, Patrizio Arrigo, and Carmelina Ruggiero, *Member, IEEE*

***Abstract*— MicroRNAs (miRNAs) are small non-coding RNAs that regulate fundamental cellular processes in diverse organisms and that have an important function in gene expression regulation. miRNAs seem capable to concurrently modulate hundreds of target genes. Their abnormal expression is emerging as important element in many pathological conditions. The identification of microRNA binding sites on those proteins that can be disease biomarker is fundamental to design synthetic artificial oligomers. In this paper we suggest a method, based on signal processing, to filter out potential miRNA recognition sites in the 3' UTR region of mRNAs.**

I. Introduction

MicroRNAs (miRNAs) are the more extensively studied non-coding RNAs. They are heavily implied in the control of many different cellular functions. The altered expression of miRNA is often associated with diseases such as, for example, cancer or diabetes. A miRNA controls gene expression, at post-transcriptional level, by interaction with specific recognition sites. These nucleotide motifs are mostly present in the 3' UTR of a messenger RNA. The identification of those miRNA that control a single protein level is important to estimate the post-transcriptional control. In both cases, from miRNA to protein and reverse, it is important to screen, in a reliable way, the recognition sites for all the potentially interacting miRNAs. Currently only a limited number of protein targets have been experimentally validated. This limitation depends on the capability of each miRNA to inhibit hundreds of different genes. The majority of computational tools [1-3] for target prediction tend to originate, taking a statistical score into account, a redundant list of potential targets. These methods generally are founded on sequence pattern matching between mature miRNA and a messenger RNA (mRNA).

Digital Signal Processing (DSP) methods were applied for a long time in biosequence analysis [4]. In many cases nucleotides have been represented by lexicographic order (A=1, C=2, G=3, T/U=4) or by different binary codes that can represent not only the nucleotide but also its properties such as number of H-bonds or weak (A, T/U) or strong (G, C). A very promising method to code nucleotide is based on electro-ion interaction potential. [5] Our analysis has been focused on identification, by signal processing methods, of potential binding sites for a specific miRNA.

II. Methods

Dataset preparation an compositional analysis

We have selected the human mir-636 because its precursor shows a very high content of GC respect the other human pre-miRNA. This compositional property seems potentially correlated with mutation propensity (CpG) [6]. The mir-636 seems to be downregulated in Cancer (tumor suppressor miRNA) [7]. We have classified the potential targets retrieved from Microcosm (http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/) [8] (830 potential different targets). This preliminary classification has allowed to filter out a limited number of potential targets with high score. Here, for sake of simplicity, we present an excerpt of statically significant proteins. In this paper we have selected a set of putative targets for mir-636. This microRNA is characterized by a high content of CpG dinucleotide.
Proteins that are been taken into account are listed in table I.

TABLE I. CONSIDERED PROTEINS

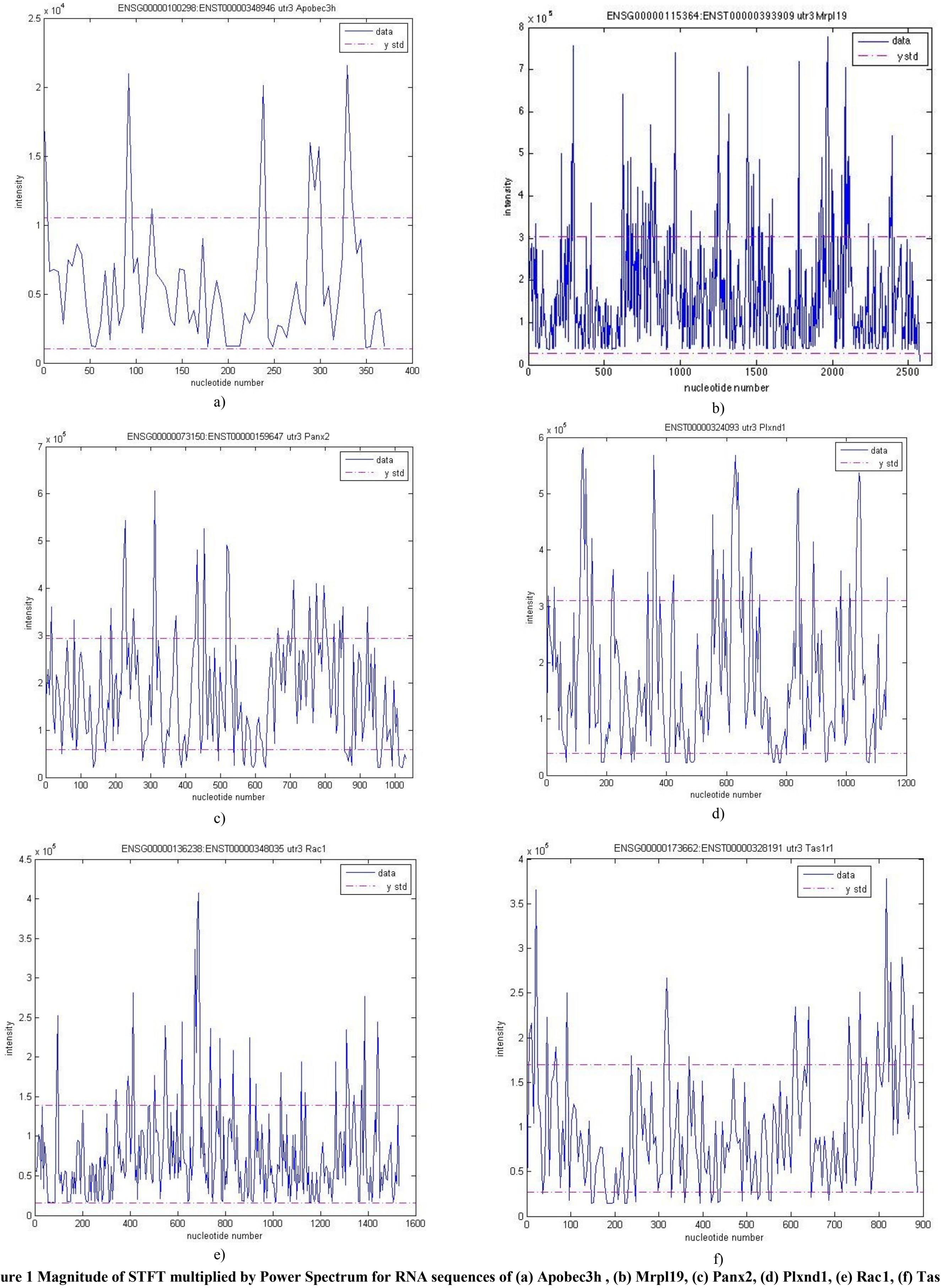
Protein	Acronym
taste receptor 1	Tas1r1
ras-related C3 botulinum toxin substrate 3	Rac3
ras-related C3 botulinum toxin substrate 1	Rac1
pannexin 2	Panx2
Mitochondrial ribosomal protein L19	Mrpl19
apolipoprotein B mRNA editing enzyme	Apobec3h
plexin D1	Plxnd1

3'UTR FASTA sequence has been retrieved from Ensembl database (www.ensembl.org). [9] Resonant Recognition Model (RRM) [10, 11] has been applied to each sequence of our dataset. RRM is a physical and mathematical model which interprets biological sequence linear information using signal analysis methods in order to treat the primary sequence as a discrete signal. In this case we have applied RRM to a RNA sequence.

Signal Processing analysis

The first step is the transformation of the symbolic sequence into a numerical one. Each nucleotide is mapped into a value

Manuscript received March 29, 2011. Patrizio Arrigo is with the CNR Institute of Macromolecular Studies (ISMAC), Section of Genoa, Via De Marini 6, 16149 Genoa, Italy (E-mail: arrigo@ge.ismac.cnr.it). All other authors are with the Department of Communication, Computer and System Sciences, Nanobiotechnology and Medical Informatics Laboratory University of Genoa, Via all'Opera Pia, 13, 16145 Genoa, Italy (*corresponding author, phone: +390103532991; fax: +390103532154; E-mail: carmel@dist.unige.it).



a)

b)

c)

d)

e)

f)

Figure 1 Magnitude of STFT multiplied by Power Spectrum for RNA sequences of (a) Apobec3h , (b) Mrpl19, (c) Panx2, (d) Plxnd1, (e) Rac1, (f) Tas1r1

of electron–ion interaction potential [12, 13] (EIIP). This parameter is an estimate of average energy states of all valence electrons in a particular nucleotide. Table II shows the values of EIIP for each nucleotide. The next steps use digital signal analysis methods applied to the obtained numerical series in order to investigate the potential correlation between spectral characteristics and biological function. In particular our analysis is concentrated on the possibility to reduce the 'redundancy' of miRNA binding sites originated by multiple sequence alignment.

In our study we have applied a Short Time Fourier Transform (STFT) based method to each 3'UTR mRNA sequence to evidence possible different function related to different position in the sequences. Using this approach it is possible to correlate the signal with specific position in the nucleotide sequence.

The STFT introduce time-localization by dividing a signal into a number of short overlapping sections using a sliding window. Fourier transform is not apply to the whole signal, but only to the section selected by the sliding window.

Thus, we have a series of frequency spectra with each spectrum corresponding to a short interval of time. The STFT is defined by

$$X_{r,k} = \sum_{m=0}^{L-1} x[m+rR]w[m]e^{-j\pi km/N} \qquad (1)$$

where L is the window length, N is the DFT length, R is the shift interval, and r and k are integers such that

$$-\infty < r < \infty \quad \text{and} \quad 0 \leq k \leq N-1$$

The time and frequency resolution depends by the length of the window. An increase in the window length enhances the frequency resolution at the expense of the time resolution while a decrease in the window length improves the time resolution at the expense of the frequency resolution.

Some proposed methods [14] in literature suggest to obtain a consensus spectrum by multiplication of the DFT coefficient of some similar sequences (eg. protein sequences of same family). Conversely, in this case we can consider only one sequence at time, then the power spectral density (PSD) has been calculated We then chose a Hanning window of length 10 samples with a superposition of 5 samples. In our analysis we have selected a window of 5 based amplitude. This dimension satisfies two main clauses: a) the resulting set of combinations is easily enumerable; b) this amplitude does not overlap with the maximal number of potentially coding triplets [14].

TABLE II. ELECTRO ION INTERACTION PSEUDO POTENTIALS OF NUCLEOTIDES

Nucleotide	EIIP values
A	0.1260
G	0.0806
C	0.1340
T	0.1335

We then obtained the STFT of the sequence. Then we evaluated the squared magnitude of the columns of the STFT multiplied by the PSD. The last step is a convenient way of emphasizing the frequency where putative miRNA binding region are located and de-emphasizing all other frequencies in the STFT.

This workflow has been adopted for each considered mRNA sequence. In figure 1 the results of the procedure are summarized for each nucleotide.

III. RESULTS AND DISCUSSION

The table III summarize the number of miRNAs predicted by Microcosm system and the number of region selected by our study. The DSP allows to filter out the binding sites with highest STFT peaks that are correlated with high intermolecular interaction propensity.

The miRNA binding sites are commonly identified by different sequence alignment methods. These tools do not give an estimate of miRNA-mRNA interaction capability. Our DSP analysis, even if in a preliminary phase, has allowed to filter out a subset of 3' UTR domains, associated with spectra peaks higher than standard deviation (std), in which fall several miRNA binding sites. The peak intensity seems to be related with the number of putative overlapping

TABLE III. NUMBER OF MIRNA INTERACTING WITH GENES AND THEIR INTERACTING REGIONS

Gene name	Transcript	No. miRNAs (Microcosm)	No. of putative interaction region
Tas1r1	**ENST00000328191**	**8**	**17**
	ENST00000333172	46	5
Rac3	ENST00000306897	62	9
Rac1	ENST00000356142	18	2
	ENST00000348035	NA	17
Panx2	ENST00000159647	29	19
Mrpl19	ENST00000358788	30	9
	ENST00000393909	NA	15
Apobec3h	ENST00000348946	46	7
Plxnd1	ENST00000324093	24	14

binding sites identified by Microcosm. Our analysis has also underlined the presence of a motif with a high peak in the boundary between coding sequence and 3' UTR. Is to be noted the difference between two transcript of Tas1R1 gene. The bolded sequence, as reported in Ensembl database, seems probably refer to a different isoform of protein. This entry shows a lower number of miRNA respects to the number of putative interaction regions. This result appears to be opposite in relation to the obtained ones with the other considered sequences. The comparison between the two forms of Tas1R1, underlined in the table below, suggest the possibility to use the proposed approach to predict potential sequencing error or differences due to genetic variability. This very preliminary information suggests us to apply this method on recognition of new cryptic regulatory signals in mRNAs. Microcosm did not used all transcripts for miRNA to detect, by alignment, miRNA's binding sites. On the contrary signal processing has identified some domains that could be related with the presence of putative miRNA's interaction motifs.

Our DSP approach has also identified a motif with minimal peak in the terminal region of 3' UTR. These preliminary results suggest the possibility to obtain a relative simple potential functional landscape for 3' or 5' UTR regulatory regions.

IV. Conclusions

Our analysis underlines the potentiality of DSP in the search of miRNA binding sites. The optimisation of miRNA binding site screening requires an integrative approach that allows to combine different computational approaches. Taking this consideration into account, the STFT method could be a promising tool that can complement alignment methods in order to reduce the binding sites redundancy of pattern matching approaches. We plan to extend the analysis of 3'UTR to other ortholog gene classes considering their significance as biomarkers for different kinds of cardiovascular diseases. In the future we also intend to apply different DSP methods, such as Wavelet and S-transform, in order to improve the capability to filter out the binding sites with the highest interaction probability.

References

[1] T. M. Witkos, E. Koscianska, and W. J. Krzyzosiak, "Practical Aspects of microRNA Target Prediction," *Curr Mol Med,* vol. 11, pp. 93-109, Mar 2011.

[2] D. Long, R. Lee, P. Williams, C. Y. Chan, V. Ambros, and Y. Ding, "Potent effect of target structure on microRNA function," *Nat Struct Mol Biol,* vol. 14, pp. 287-94, Apr 2007.

[3] D. Long, C. Y. Chan, and Y. Ding, "Analysis of microRNA-target interactions by a target structure based hybridization model," *Pac Symp Biocomput,* pp. 64-74, 2008.

[4] S. A. Marhon and S. C. Kremer, "Gene prediction based on DNA spectral analysis: a literature review," *J Comput Biol,* vol. 18, pp. 639-76, Apr 2011.

[5] A. S. Nair and S. P. Sreenadhan, "A coding measure scheme employing electron-ion interaction pseudopotential (EIIP)," *Bioinformation,* vol. 1, pp. 197-202, 2006.

[6] F. Giuliano, P. Arrigo, F. Scalia, P. P. Cardo, and G. Damiani, "Potentially functional regions of nucleic acids recognized by a Kohonen's self-organizing map," *Comput Appl Biosci,* vol. 9, pp. 687-93, Dec 1993.

[7] I. Van der Auwera, R. Limame, P. van Dam, P. B. Vermeulen, L. Y. Dirix, and S. J. Van Laere, "Integrated miRNA and mRNA expression profiling of the inflammatory breast cancer subtype," *Br J Cancer,* vol. 103, pp. 532-41, Aug 10 2010.

[8] S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright, "miRBase: tools for microRNA genomics," *Nucleic Acids Res,* vol. 36, pp. D154-8, Jan 2008.

[9] T. J. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle, and P. Flicek, "Ensembl 2009," *Nucleic Acids Res,* vol. 37, pp. D690-7, Jan 2009.

[10] I. Cosic, "Macromolecular bioactivity: is it resonant interaction between macromolecules?--Theory and applications," *IEEE Trans Biomed Eng,* vol. 41, pp. 1101-14, Dec 1994.

[11] E. Pirogova, G. P. Simon, and I. Cosic, "Investigation of the applicability of dielectric relaxation properties of amino acid solutions within the resonant recognition model," *IEEE Trans Nanobioscience,* vol. 2, pp. 63-9, Jun 2003.

[12] V. Veljković and I. Slavić, "Simple General-Model Pseudopotential," *Physical Review Letters,* vol. 29, pp. 105-107, 1972.

[13] S. Glisic, P. Arrigo, D. Alavantic, V. Perovic, J. Prljic, and N. Veljkovic, "Lipoprotein lipase: A bioinformatics criterion for assessment of mutations as a risk factor for cardiovascular disease," *Proteins,* vol. 70, pp. 855-62, Feb 15 2008.

[14] P. Ramachandran, A. Antoniou, and P. P. Vaidyanathan, "Identification and location of hot spots in proteins using the short-time discrete Fourier transform," 2004, pp. 1656- 1660 Vol.2-1656- 1660 Vol.2.