New Approach to Predict microRNA Gene by using data Compression technique

Dae-Won Kim[1,3,*] Joshua SungWoo Yang[1,4,*] Pan-Jun Kim[2] In-Sun Chu[1,4], Hawoong Jeong[2], Hong-Seog Park[1,3]
[1] Department of Functional Genomics, University of Science and Technology, Daejeon, Korea
[2] Department of Physics, Korea Advanced Institute of Science and Technology, Daejon, Korea
[3] Genome Research Center, Korea Research Institutes of Bioscience & Biotechnology, Daejeon, Korea
[4] National Genome Information Center, Korean Research Institutes of Bioscience & Biotechnology, Daejeon, Korea
[*] These authors contributed equally to this work
Email : hspark@kribb.re.kr

ABSTRACT: Over the past few years, the complex and subtle roles of microRNA (miRNA) in gene regulation have been increasingly appreciated. Computational approaches have played one of important roles in identifying miRNAs from plant and animals, as well as in predicting their putative gene target. We present a new approach of comprehensive analysis of the evolutionarily conserved element scores and applied data compression technique to detect putative miRNA genes. We used the evolutionarily conserved elements [19] (see more detail on method and material) to calculate for base-by-base along the candidate pre-miRNA gene region by detecting common conserved pattern from target sequence. We applied the data compression technique [20] to detect unknown miRNA genes. This zipping method devises, without loss of generality with respect to the nature of the character strings, a method to measure the similarity between the strings under consideration [20]. Our experience to using our new computational method for detecting miRNA gene identification (or miRNA gene prediction) has been stratified and we were able to find 28 putative miRNA genes.

1 INTRODUCION

Several hundred novel genes encoding transcripts containing short-strand RNA hairpins, microRNA (miRNA) genes are genes for which the transcribed RNA is either by arresting the translation of messenger RNAs (mRNA) or by the cleavage of mRNA [1,4]. Instead, the RNA transcript is the functional end product rather than an intermediary messenger that represents a class of regulatory small non-coding RNAs of ~19-24 nucleotides though base pairing to partially complementary sites in the 3' untranslated regions (UTRs) across a brad range of metazoan from plant and human [1-3]. Recent study [17] has been suggested to examine not only 3' UTR sequences but also 5' UTRs and the result of Bradely and colleagues' study [18] discovered ~70% of mammalian miRNA genes are located in defined transcription units, which are found in the introns and integenic regions. This fact is also support by review paper from Kim [28] that suggests three bases of miRNA genomic locations: exonic miRNA in non-coding transcription units, intergenic miRNA in non-coding transcription unit, and intergenic miRNA in protein-coding transcription units.

Performing direct cloning experiments is enabled to detect many miRNAs; however, significant variation in their expressions is very difficult to clone low abundance miRNAs [8] such as Dicer binding motifs [13], Drosha recognition motif [14], and miRNP complex protein binding motifs [15]. For those problems, recent research study has been using the computational approaches to predict miRNAs in genome sequence data with reasonable efficiency [2, 3, 7-12] and detect unknown miRNAs which are not screened with experimental method.

The large majority of these methods are heavily relies on the function characterization to perform the nature of pattern recognition between miRNAs and target genes in vertebrate and plant. General computational tools predicted close homologs of target gene with statistically conserved patterns of known miRNA. In plants, miRNAs are nearly perfectly complementary to their target gene, the computational approach are very straightforward and powerful. However, in vertebrate, miRNAs target is very difficult to detect because miRNAs target pairings are not entirely complementary and there are lack information of known miRNAs, which are generalized for genome-wide searches. The principles behind these computational approaches are this following: the conservation of target 3' UTR sequences in orthologous genes and the kinetics and thermodynamics of the association between the miRNA and its target, as determined by RNA folding programs [16]. Many of them are frequently fail to detect miRNAs that lack detectable homologs.

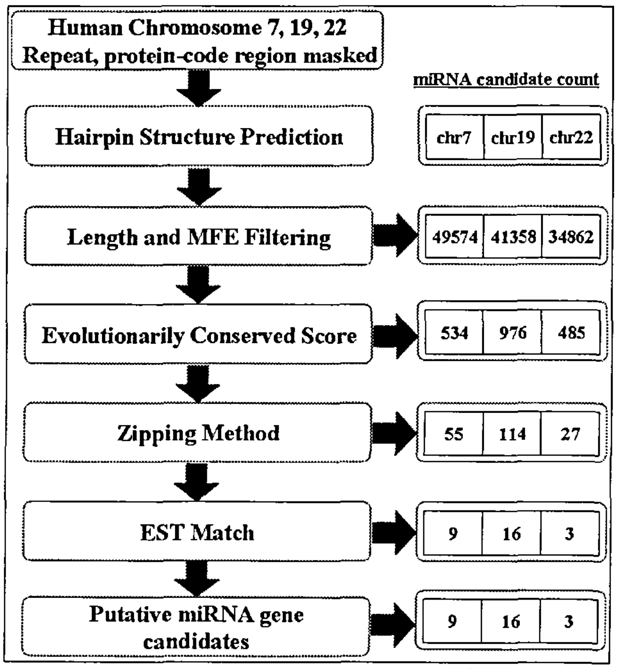

Human Chromosome 7, 19, 22
Repeat, protein-code region masked
miRNA candidate count
Hairpin Structure Prediction
chr7
chr19
chr22
Length and MFE Filtering
49574
41358
34862
Evolutionarily Conserved Score
534
976
485
Zipping Method
55
114
27
EST Match
9
16
3
Putative miRNA gene candidates
9
16
3

Figure 1. Computational predicted miRNA pipeline. This is flow chart of predicted miRNA pipeline. We used this following

order: masked sequences selection (without repeat, protein-coding region), hairpin (stem-loop structure) with calculation of MFE, Evolutionarily conserved score, zipping method, and EST match.

The purpose of the current investigation is based on fact of Bradeyl and colleagues' study [18] and Kim's [28] review paper, we developed new pipeline of predicted miRNA to detect new specific miRNA target on intron and intergenic regions on human chromosome 7, 19, and 22. This pipeline is an integrated system that including this following (shown in figure 1): combination of thermodynamics-based on secondary structure of RNA with minimum free energy measurement, evolutionarily conserved element scores measurement [19], zipping score (based on application of zipping theory [20]), and *Homo sapiens* EST pattern recognition. We were able to identify 28 miRNA genes on chromosome 7, 19, and 22 of *Homo sapiens.*

2 Method and Material

2.1 *Home sapiens* genome sequences and miRNA reference dataset

For target of *Homo sapiens* genome sequence dataset, sequence and annotation were downloaded the masked version of the genome sequences on *Homo sapiens* chromosome 7, 19, and 22 from the Ensemble dataset (*Homo sapiens* release NCBI build 35, April 2005, ftp://ftp.ensemble.org/pub/current_ji.am/data/fasta/dna). Those masked version of the genome sequences are interspersed repeats and low complexity regions to detect with the repeatmasker tool also masked by replacing repeats with 'N's. Using of EnsMart [21], we found the position of known protein-encoding genes and pseudogens region of target chromosomes to mask by replacing those with 'N'. For miRNA reference dataset, hairpin all mammalian miRNA sequences (*C. familiaris, D. rerio, G. Gallus, H. sapiens, M. muscules, and R. novegicus*) were obtained from the Rfam miRNA registry release 6.0 (http://www.sanger.ac.uk/Software/Rfam/mirna/) [22].

2.2 Identifying hairpin structure for candidates miRNA

Method to predict secondary structure from the prime sequence information is based on minimizing the free energy of the molecule by maximizing the number of favorable base pairing interactions [23]. Therefore we fold the entire target human genome sequence using the Vienna package [25] in windows of 100 nucleotides with an overlap of 80 nucleotides. All hairpin structures that have at least 60 nucleotides long were extracted from the minimum free energy (MFE) fold by below -30 kcal / mol^{-1}. We choose – 30 kcal / mol^{-1} as cut off point because this is the standard deviation of reference miRNA datasets' MFE.

2.2 Evolutionarily conserved element scores measurement

The multiple alignments, predicted conserved elements in vertebrate genomes (using genome-wide multiple alignments of five vertebrate species (human, mouse, rat, chicken, and Fugu rubripes)), and base-by-base conservation scores, known as PhastCons [19], were download from UCSC genome browser (http://www.cse.ucsc.edu/~acs/conservation). This conservation pattern is especially impressive in the clusters. We measured all of candidate miRNAs and references miRNA for conserved pattern element scores then used the mean values of reference miRNA conserved pattern element scores (0.8) for cut off point of pre-miRNAs selection. After this filtering, using 15 base of window shift size to scans all of selection of pre-miRNA selection which pass by average of miRNA conserved pattern element scores and reference miRNAs to measure the high peak value of gene regions. We use a cut off point as the mean of reference miRNAs (0.95) for scanning of 15 bases of window shift size filter out intermediate-miRNA genes selection.

2.3 Zipping Algorithms

To apply zipping method [20], based on data-compression technique, is the computational measurement of remoteness of two different sequences which are reference miRNA and intermediate-miRNA sequence. This compression algorithm is very powerful tool for the measure of the relative entropy by zipping two sequences into one. We used this method to calculate the remoteness of two different sequences to select high quality of predicted miRNA candidates by modified equation of the language tree and zip algorithm [20]. In order to measure the remoteness of two different sequences (reference miRNA (A) and intermediate-miRNA (B)), we create a new sequence A + B by simply appending B after A. The length of A+B, L_{A+B}, now zipped, for example using gzip, and the measure of the length of B (L_B) in the coding optimized for A will be, $(L_{A+B}) - L_A$, where indicated the length in bits of the zipped file and L_A is the length of zipped A. The divergence D_{AB} per character between A and B will be estimated by

$$D_{AB} = \frac{L_{A+B} - L_A}{|B|}$$

where |B| is the number of characters of the sequence of B.

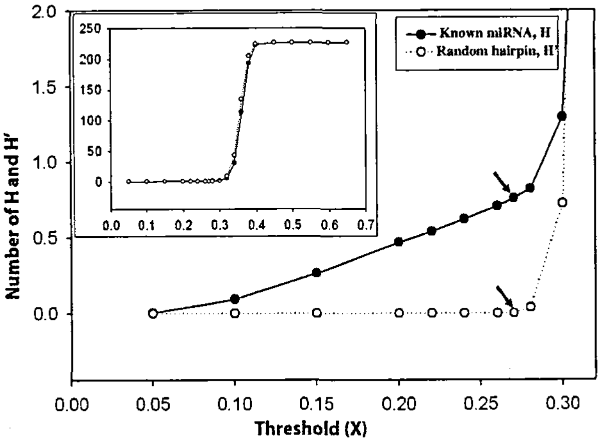


Figure 2. Estimate of the number of coordinately regulated targets for set of miRNA candidates by zipping method. The graph indicates the procedure of choosing the threshold of divergence of reference miRNA genes dataset and miRNA candidate genes. Using of two arrows is a point at the location of threshold and we use sub-small window to show the overall view of plot.

To eliminate false positive of predicted miRNAs, we defined the threshold of predict miRNA selection. For selecting of threshold, we create a training set of reference miRNAs to measure individual divergence (or remoteness), $D_{AH,}$ of two different sequences (reference miRNA gene datasets (A) and 227 known human miRNA (H) [22]) without human miRNAs from reference miRNA gene dataset. After measurement of D_{AH}, we perform this following to select the threshold;

1) For given sequence of 227 known human miRNA [22], H, evaluate the corresponding D_{AH}, and take the means of whole reference miRNA gene sequence, $\overline{D}_H$. For arbitrary value X, we obtain the number N(X) of sequence H which satisfies $\overline{D}_H \leq$X, and plot N(X) as a function of X.
2) We randomly select 227 hairpin structure results, H'; we evaluate similar way in step 1. For an arbitrary value X, we obtain the number N'(X) of H' which satisfy $\overline{D}_{H'} \leq$ X, and plot N'(X) as a function of X.
3) As illustrated in figure 2, N'(X) is close to zero if $X \leq 0.27$, and in the same domain, N(X) has the value distinctively larger than N'(X). When 0.27 < X, N(X) and N'(X) increase sharply, and become almost saturated at maximum of 227. In other words, (a) for 0.27 < X, most of the cases for H and H' satisfy $\overline{D}_H \leq$X, $\overline{D}_{H'} \leq$X, while (b) for $X \leq 0.27$, $\overline{D}_H$ of H clearly carries the significance in the similarity between the reference miRNA and H, contrary to the case with $\overline{D}_{H'}$ of randomly selected sequence. Consequently, the less than 0.27 D_{th}, the threshold of divergence, is chosen as, the improved the selectivity of actual miRNA precursors among the random sequence.
4) By assuming D_{th} as 0.27, we identify the number of H_S satisfying $D_{AH} \leq D_{th}$, for at least one of reference sequence, A, then obtain 204 of the total 227 as the potential candidates of actual miRNA precursors. Consequently, assuming D_{th} as 0.27 does not scarify the sensitivity at all for the true-positive detection on actual miRNA precursors as well as the high selectivity revealed in step 3.

After this following 4 steps, we select the threshold of divergence, D_{th} ,as 0.27 (shown in figure 2).

2.4 EST pattern recognition

The subgroup of *Homo sapiens* of the publicly available EST databases (March, 2005) was searched using DeCyher BLASTn(http://www.timelogic.com/decypher_citations.html) [24] by comparing all ESTs to all previously result of data compression algorithms. DeCyher BLASTn parameters were setting by default with cut off of E-value is 1.0^{-4}. The candidates satisfied our requirements (expect coverage of miRNA genes were 100%; and alignment were 100 %) then we select this result as putative miRNA gene candidates.

3 Result

Using of the Vienna package [25], it helps to estimate fold the RNA secondary structure of all of masked target human sequences and reference miRNA (also known as known miRNA) genes dataset (see more detail on method and material) and calculate the minimum free energy. The result of Vienna package [25], we used the extracting method of stem-loops structures on chromosome 7, 19, and 21 with maximum -30 kcal / mol^{-1}. We selected pre-miRNA genes this following (shown in figure 1): 49,574 for chromosome 7, 41, 358 for chromosome 19, and 34,864 for chromosome 22.

Using of evolutionarily conserved element scores by base-by-base, we can estimate the conservation pattern where typical peaks of high conservation are found in close proximity (shown in figure 3). The conservation peaks span the miRNA and its precursor. We find similar behavior with Altuvia and colleague study [9] that in many case there is a trough in the middle of the conservation peak, generating saddle-like shape, which results from the lower conservation of the loop region in the secondary structure of the miRNA precursor (shown in figure 3). The result of highly strict conserved element score examination (see more detail on method and material), we selected 534 of 49,574 miRNA in chromosome 7, 976 of 41,358 miRNA in chromosome 19, and 485 of 34,862 miRNA in chromosome 22 (shown in figure 1).

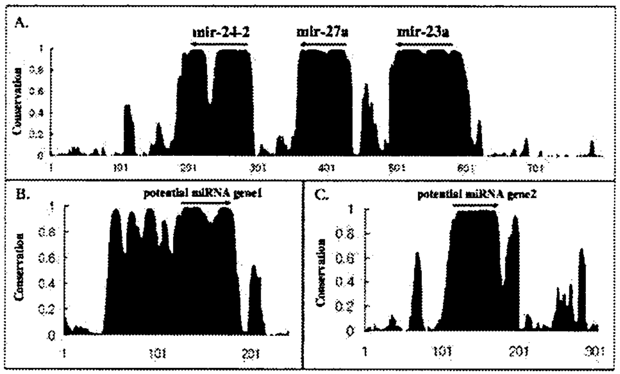
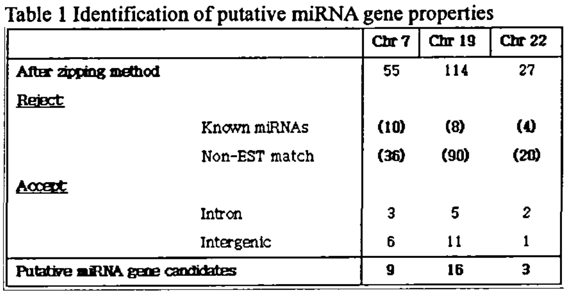


Figure 3. Evolutionarily conservation patterns of known and predicted *Homo sapiens* chromosome 19 miRNA genes. The y axis represents multiple alignment PhastCons score [19] and the *x*-axis displays the relative positions of miRNA genes. (A) is showing the known miRNA gene of high peak conserved pattern regions. Known miRNAs are designated by their Rfam name omitting the 'hsa' prefix. Cluster on chromosome 19 is located at relative position 13807900-13808700. (B) and (C) are indicating predicted potential miRNA genes, which are relative position.

In figure 2 illustrated a graph of the output of zipping method which helps us to determine the appropriate threshold of the divergence D_{AB} , known as the score of remoteness between reference miRNA gene dataset and intermediate miRNA genes (more detail on method and material). This graph indicates that less than 0.27 of threshold, known as divergence of threshold, is chosen as, the improved the selectivity of actual miRNA precursors among the random sequence and does not scarify the sensitivity at all for the true-positive detection. Output indicates that we collected 55, 114, and 27 of chromosome 7, 19, and 22. We used the final silico confirm as the EST matches.

We only used the prefect match with EST, it provide very reasonable output (shown in Table 1). We found 9, 16, and 3 putative miRNA genes of chromosome 7, 19, and 22. We annotated those putative miRNA genes into target chromosomes; they are located on intron and intergenic regions.

Table 1 Identification of putative miRNA gene properties



		Chr 7	Chr 19	Chr 22
After zipping method		55	114	27
Reject				
	Known miRNAs	(10)	(8)	(4)
	Non-EST match	(36)	(90)	(20)
Accept				
	Intron	3	5	2
	Intergenic	6	11	1
Putative miRNA gene candidates		9	16	3

4 Discussion

Many studies that verify in silico predictions with wet-bench experiment will crucial to the testing and refinement of miRNA computational databases and algorithms [16]. The precise rules and energetic for pairing between a miRNA and its mRNA target genes still have remain in a lack of information and it is very difficulty challenges in computational predictions. Therefore, many computational approaches for the identification of potential miRNA targets site are at risk of having a substantial rate of false positives and false negative [27]. This means, the statistical information of miRNA genes is insufficient to identify miRNA genes, which makes it difficult to predict [3]. Though this difficultly, numerous studies suggest two approaches; defined conversed region by comparative approach; and defined common secondary structure of putative miRNA gene candidates.

The detecting common secondary structure approach is a straight-forward for identification of new targets such as gene, regulatory motif, protein, RNA and chemical [18]. We used the Vienna package [25] to estimate folded the thermodynamics base on secondary structure for candidates with measurement of minimum of free energy, known as MFE. MFE is estimated by considering the minimum energy of successive base pair or increase by the destabilizing energy associated with non-complementary bases [23] and check the stability of secondary structure. For this reason, it is very important for the predicted MFE to select miRNA candidates from hairpin structure. The process of this examination, we discovered that known miRNA (has-miRna-330, has-miRna-25, and has-miRna-7b) did not satisfy our requirements. This output is remained as unknown. We assume that it might be prediction error from pervious study or Vienna package might cause the technical error.

The comparative analysis approach emphasizes the importance of multi-organism comparison to detect the conserved regions for putative miRNAs. Because of miRNAs are highly conserved across different organism, all previously reports [29, 30]. In this investigation, we used the evolutionarily conserved elements [19] (see more detail on material and method) to calculated for base-by-base along the candidate pre-miRNA gene region. It was easy derive the conservation patterns of reference miRNAs and the proximal regions. This result supported to Altuvia et al and colleagues study [31] that miRNA gene might be the evolutionary and functional implications.

This investigation focused on a concept of remoteness (or similarity) measurement between pairs of sequences based on their relative informatics content. From our analysis of conserved pattern region, known miRNA (has-miRna-371, has-miRna-373, has-miRna-372, has-miRna-148a, and has-miRna-196b) were failed to receive enough scores of conserved pattern element. Because we set to be highly strict cut off to reduce false positives to not collect non-conserved region of candidate.

We applied the zipping method [20] into computational predicted miRNA pipeline (shown in figure 1). This zipping method devises, without loss of generality with respect to the nature of the character strings, a method to measure the similarity between the strings under consideration [20]. This method is highly versatile and general that it does not require *a priori* knowledge about the statistics of the investigated data. This method is very powerful tool, however; it has not been clearly verified whether the zipping method outperforms other methods including Markov Chain approach or not. Our experience to using this computational approach for detecting miRNA gene identification (or miRNA gene prediction) indicated close to meets our expectation, however, there is a problem slow performing of computational time. Because we did not confirm by experimental approach, our prediction is remained questionable for quality of output.

Though this problem, we tried to use the EST datasets to measure the quality of output; however, it did not help. We are still in progress to confirm our result by experiment approach. Recent study [32] discovered that miRNA genes are located on repeat sequences in mammalian. The coverage of repeat sequences on human genome is 41%; therefore, many unknown miRNAs located in repeat sequence regions. However, the given of target human genome sequences (chromosome 7, 19, and 22) in current investigation, we used mask repeat and protein-coding region of target sequence that it was out of rang for our investigation for detecting the repeat sequence region in miRNA genes. In future study, we are planning to scan whole genome of *Homo sapiens* on intron and intergenic regions with experimental confirmation. We hope that refinements of the computational pipeline, especially zipping method approach, used here for the predication of miRNA gene may cope with this difficulty in the future.

5 Acknowledgement

This research was supported by a grant from *KRIBB Research Initiative Program.* P.K. and H.J. acknowledge financial support from the Ministry of Science and Technology through Korean Systems Biology Research Grant (M10309020000-03B5002-00000).

REFERENCES

[1] Bartel, D.P. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, 116, 281 – 297, 2004

[2] Klein, R.J. Finding noncoding RNA genes in genomics sequence. PHD Thesis August, 2003

[3] Nam, J.W. Human microRNA prediction though a

probabilistic co-learning model of sequence and structure. Nucleic Acids Res., 33, 3570-3581, 2005

[4] Mallory A.C. and Vaucheret H. MicroRNAs: something important between the genes. Curr. Opin. Plant Biol. 7, 120-125, 2004

[5] Bonnet E. Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. Proc. Natl. Acad. Sci USA 101, 1511-11516.

[6] Jing Q et al. Involvement of MicroRNA in AU-Rich Element-Mediated mRNA instability. Cell 120, 623-634, 2005

[7] Ambros, V et al. A uniform system for micoRNA annotation. RNA 9:277-279, 2003

[8] Adai, A et al. Computational prediction of miRNAs in Arabidopsis thaliana. Genome Res. 15:78-91, 2005

[9] Altuvia, Y et al. Clustering and conservation patterns of human microRNAs. Nucleic Acids Res., 33, 2697-2706, 2005

[10] Wang, X.J. et al. Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. Genome Biol. 5:R65 2004

[11] Rajewsky, N and Socci, N.D. Computational identification of micoRNA targets. Devel. Biol., 267:529-535 2003

[12] John B et al. Human MicroRNA Targets Plos. Biol. 2(11):e363 2004

[13] Song et al. The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effectors complex. Nat. Struct. Biol., 10(12):1026-32, 2003

[14] Lee Y et al. The nuclear Rnase III Drosha initiates microRNA processing. Nature, 425:415-419, 2003

[15] Dostie J et al. Numerous microRNPs in neuronal cells containing novel microRNAs. RNA, 9:180-186, 2003

[16] Brown J.R and Sanseau P. A Computational view of microRNAs and their targets. Drug discv. today: biosilico, 10(8), 595-601. 2005

[17] Ambros, V. The functions of animal microRNAs Nature 431, 356-363. 2004

[18] Griffiths-Jones et al. Identificaiton of mammalian microRNA host genes and transcription units. Genome Res. 14, 1902-1910, 2004

[19] Siepel A. et al. Evolutionrily conserved elements in vertebrate, insect, worm, and yeast genome. Genome Res. 1-17

[20] Benedetto D. et al. Language Trees and Zipping. Phys. Rev. Lett. 88(4), 1-4

[21] Kasprzyk A et al. EnsMart: A generic system for fast and flexible access to biological data. Genome Res. 14: 160-169 2004

[22] Griffiths-Jones S. The microRNA Registry. Nucleic Acids Res., 32, 109-111, 2004

[23] Zuker, M and Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res., 9, 133-148, 1981

[24] Margulies E.M et al. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc. Natl. Acad. Sci USA 102(13), 4795-4800 2005

[25] Hofacker, I.L. et al. Fast Folding and comparison of RNA secondary structures. Monatch. Chem 125:167-188, 1994

[26] Lempel A. and Ziv. J. IEEE Trans. Inf. Th., 337-343, May 1977

[27] Enright A. J. and et al. MicoRNA targets in Drosophila. Genome Biol. 5(I):RI

[28] Kim, V.N. MicroRNA Biogenesis: coordinated cropping and dicing Nature 6, 356-385, 2005

[29] Lim, L.P. et al. Vertebrate microRNA. Science, 299:1540

[30] Lai. E.C. et al. Computational identification of Droshophila microRNA genes. Genome Biol., 4:R42.

[31] Altuvia, Y et al. Clustering and conservation patterns of human microRNAs. Nucleic Acids Res., 33(8), 2697 – 2706. 2005

[32] Smalheise N.R. and Torvik V.I. Mammalian microRNAs derived from genomics repeats. TRENDS in genetic 21(6):322-326, Jan 2005