Finding conserved well-ordered RNA structures in genomic sequences

Shu-Yun Le
Laboratory of Experimental and Computational Biology
NCI Center for Cancer Research
National Cancer Institute, NIH,
Bldg 469, Room 151, Frederick, Maryland 21702

Jacob V. Maizel, Jr.
Laboratory of Experimental and Computational Biology
NCI Center for Cancer Research
National Cancer Institute, NIH,
Bldg 469, Room 151, Frederick, Maryland 21702

Kaizhong Zhang
Department of Computer Science
University of Western Ontario
London, Ontario, N6A 5B7, Canada

Recent advances in RNA studies show that the well-ordered, structured RNAs perform a broad functions in various biological mechanisms. Included among these functions are regulations of gene expression at multiple levels by diversified ribozymes and various RNA regulatory elements. The discovered microRNAs (miRNAs) with a distinct stem-loops are a new class of RNA regulatory elements. The prediction of those well-ordered folding sequences (WFS) associated with the RNA regulatory elements in genomic sequences is very helpful for our understandings of RNA-based gene regulations. We present here a new computational method in searching for the conserved WFS in genomes. In the method, the WFS is assessed by a quantitative measure E_{diff} that is defined as the difference of free energies between the computed optimal structure (OS) and its corresponding optimal restrained structure where all the previous base pairings in the OS are forbidden. From those WFS with high E_{diff} scores, the conserved WFS is determined by computing the maximal similarity score (MSS) between the two compared structures. In practice, we first search for those distinct WFS with high statistical significance in genomic sequences and then seek for those conserved WFS with high MSS. The potential and implications of our discoveries in the genome of *Caenorhabditis elegans* are discussed.

Keywords: microRNA; well-ordered folding sequence; structural similarity.

1. Introduction

RNA is a conformationally polymorphic macromolecule and is synthesized from the DNA template in transcription. Though RNA is synthesized in single-stranded chain, almost every RNA molecule has structure that includes various double helical regions of the base pair formed by itself in the correct antiparallel orientation between complementary segments. In addition to Watson-Crick A:U and G:C base pairs, wobble G:U and other non-canonical base pairs also contribute to the structural constraints in the secondary and tertiary structure of RNA molecules.

Recent advances in studies of non-coding RNAs (ncRNAs) and RNA interference (RNAi) indicate that RNA is more than a messenger between genome and protein. The ncRNAs are involved in various regulatory mechanisms of gene expression at multiple levels [1,2,3,4]. The well documented includes transcriptional mediation, RNA processing and modification, mRNA stability and localization, and the translation of mRNA into protein [5,6,7,8,9,10,11]. The functional RNAs that can confer the regulatory activity comprise transfer RNA, ribosomal RNAs, self-cleavage ribozymes [5], small microRNAs (miRNAs) [4,7] and various RNA regulatory elements, such as iron-responsive element (IRE) in the non-coding region (NCR) of ferritin mRNAs [6], internal ribosome entry sequence in the 5' NCR [10] and Rev response element in HIV [8]. It has been suggested that the functional RNAs possess well ordered conformations that are both thermodynamically stable and uniquely folded [12,13]. Those functional structured RNAs (FSR) almost always have conserved structural motifs manifested in the specific combinations of base pairings and distinct loop sequences in the folded stem-loops. It is the conserved, structural feature formed uniquely in the FSR that plays a crucial role in the regulatory mechanism.

The newly discovered miRNAs of about 22 nucleotides (nt) can control developmental timing in *Caenorhabditis elegans* (*C.elegans*) and repress the translation of their target genes by binding to the 3' untranslated regions of the mRNAs [7]. The RNA silencing underlies several important and highly related gene regulatory mechanisms [3]. It is interesting to note that most of these miRNAs have phylogenetically conserved sequences of about 22 nt and their RNA precursors are of about 80 nt in length. The precursors can form a conserved fold-back stem-loop structure across the divergent species in which the conserved sequence with about 22 nt is within one arm containing at least 16 base-pairings [14,15]. It has been suggested that a large number of miRNAs (up to hundreds) may be encoded per genome [15]. Moreover, genome analysis and comparison indicated that about 98% of the transcriptional output of human genome is ncRNAs [16]. Knowledge discovery from such ncRNAs in genomic sequences by computational method is highly desirable.

The complete genome sequences of human, rat, *C.elegans* and other pathogenic bacteria provide the fundamental information useful for us to explore biological properties. Analysis of the massive sequence data requires sophisticated bioinformatics tools. Recently, we have developed a new method of discovering well-ordered folding sequences (WFS) in a genome [17,18] and of computing quantitatively the maximal similarity between two RNA structures [19]. With the combination of the two algorithms we develop a new procedure in searching for distinct fold-back structures that are expected to be related to the known miRNAs in genomes. In this study, we first search for WFS with high statistical significance by scanning successive segments of both 70 and 80 nt along the *C.elegans* genome. We collect the distinct WFS sequences and their RNA secondary structures. We then compare each of these

RNA structures with that of the known miRNAs. Those distinct stem-loops with high similarity to the structural morphologies of known miRNAs are determined. The common structural motifs can be revealed.

2. Methods

2.1. *Search for distinct WFS*

To search for functional RNA elements with structure dependent functions in a genome sequence we need a quantitative measure to characterize the thermodynamic stability and well-ordered conformation of a RNA structure folded by a segment sequence. For an arbitrary RNA segment s, we define $E_{diff}(s)$ as the quantitative measure, $E_{diff}(s) = E_f(s) - E(s)$, where $E(s)$ is the minimal energy of the optimal folded structure (minimal energy structure) on s and $E_f(s)$ is the minimal energy of a constrained optimal structure on s where all previous base pairings in the optimal structure are prohibited. The measure $E_{diff}(s)$ can signify how thermodynamic stable and well-ordered the predicted RNA secondary structure of the segment is. The program EDscan [17] adapt the dynamic programming algorithm and Turner energy rules [20,21] to compute the minimal energies of RNA foldings. The search for WFS in a sequence is done by scanning successive segments along the sequence to calculate a standardized z-score ($Zscr_e$) for each of the successive segments. Here we define $Zscr_e(s)$ as, $Zscr_e(s) = (E_{diff}(s) - \bar{E}_{diff})/std_{diff}$
where $\bar{E}_{diff}$ and std_{diff} are the sample mean and standard deviation of $E_{diff}(s)$ computed from those successive segments. In this study we slide both a window of 70-nt stepped with 3-nt each time and a window of 80-nt stepped with 5-nt each time along the *C. eleg* genome. Those segments with $Zscr_e(s) > Zscr_e^0$ are selected to be candidates of WFS in the genome sequence.

The value of $Zscr_e^0$ is often determined using simple Monte Carlo simulations. In the simulation, we first select a natural sequence of about 2500-nt that includes a specific miRNA precursor. We compute the $Zscr_e(s)$ distribution of all 80-nt segments in the real sequence by EDscan. We then generate a set of 50 random sequences by randomly shuffling the natural sequence. With the same parameters as used in the computation of the real sequence we repeatedly compute $Zscr_e(s)$ distribution in the 50 random sequences by EDscan. With the comparison of the $Zscr_e(s)$ distributions between natural and randomly shuffled sequences, we can define a reasonable threshold of $Zscr_e^0$ for predicting the potential WFS in a genome sequence. The folded structures of those detected potential WFS are then computed by mfold [21] and used in the following structural comparison.

2.2. *RNA structural comparison*

Following the tradition in sequence comparison, we define three basic edit operations, relabel, delete, and insert, on a RNA structure. In the program rna_match [19], each operation can be applied to either a base pair or an unpaired base. With score functions associated with the edit operations for both unpaired bases and base pairings we can compute a maximal similarity score (MSS) between two RNA structures using the optimal number of weighted operations.

The dynamic programming algorithm [17,22,23] of structure comparison used in rna_match is briefly described here. Let $R_1[1\cdots m]$ and $R_2[1\cdots n]$ be the two given RNA structures. $M(i_1, i_2\ ;\ j_1, j_2)$ is used to represent MSS between the two substructures $R[i_1\cdots i_2]$ and $R_2[j_1\cdots j_2]$. Suppose that we want to compute the MSS between $R_1[1\cdots i]$ and $R_2[1\cdots j]$. If both $r_1[i]$ and $r_2[j]$ are unpaired bases, then we have

$$M(1,i\ ;\ 1,j) = \max \begin{cases} M(1,i-1\ ;\ 1,j) + del(r_1[i]) \\ M(1,i\ ;\ 1,j-1) + ins(r_2[j]) \\ M(1,i-1\ ;\ 1,j-1) + rel(r_1[i], r_2[j]) \end{cases}$$

Where del(r_1[i]), ins(r_2[j]) and sub(r_1[i], r_2[j]) are cost scores associated with the operations of deletion, insertion and relabel of unpaired bases, respectively.

If $i' < i$ and $(r_1[i'], r_1[i])$ is a base pair and $j' < j$ and $(r_2[j'], r_2[j])$ is a base pair, then we have the following if in $M(1, i-1\ ;\ 1, j)$ $r_1[i']$ is deleted and in $M(1, i\ ;\ 1, j-1)$ $r_2[j']$ is inserted.

$$M(1,i\ ;\ 1,j) = \max \begin{cases} M(1,i-1\ ;\ 1,j) + del((r_1[i'], r_1[i])) \\ M(1,i\ ;\ 1,j-1) + ins((r_2[j'], r_2[j])) \\ M(1,i'-1\ ;\ 1,j'-1) + M(i'+1,i-1\ ;\ j'+1,j-1) \\ \qquad +rel((r_1[i'], r_1[i]), (r_2[j'], r_2[j])) \end{cases}$$

Thus, we can compute MSS between two structures of R_1 and R_2 by a dynamic programming algorithm. We consider the smaller substructures first and eventually consider the whole structures R_1 and R_2.

The MSS score will depend on indel scores and the substitution matrix. Currently these values are determined by heuristics. In the future, with a large collection of RNA structural data, these can be determined by statistics.

3. Results and Discussion

3.1. *miRNAs and the corresponding WFS discovered in C.elegans*

C.elegans genome includes chromosomes I-V and chromosome X. Their lengths are from 16 to 21.3 million nts. For each chromosome sequence we computed $Zscr_e$ distribution by sliding a 80-nt window stepped with 5-nt each time along the sequence. We also computed $Zscr_e$ distribution by sliding a 70-nt window stepped with 3-nt each time along these sequences. We found that most of known miRNAs encoded in *C.elegans* were coincident with statistically significant WFS (see Table 1). The miRNAs of $21-24$ nt are located at either the right or the left arm of the folded stem-loops of the corresponding WFS summarized in Table 1.

Table 1. Known miRNAs and the corresponding WFS identified in *C.elegans* Genome.

Gene	Chromosome and Location			WFS	Zscr
lin-4	II	5902232-5902252		5902221-5902300	4.74
let-7	X	14743369-14743390	(-)	14743402-14743323	7.11
mir-1	I	4514825-4514845	(-)	4514814-4514893	7.73
mir-2	I	7697273-7697295	(-)	7697337-7697268	4.12
mir-34	X	2708647-2708668	(-)	2708601-2708680	8.27
mir-35	II	11537564-11537585		11537516-11537595	9.18
mir-36	II	11537669-11537690		11537616-11537700	3.59
mir-37	II	11537789-11537810		11537741-11537820	5.93
mir-38	II	11537886-11537907		11537836-11537915	8.65
mir-39	II	11538039-11538060		11537991-11538070	7.58
mir-40	II	11538135-11538156		11538086-11538165	9.31
mir-41	II	11538264-11538285		11538211-11538290	3.83
mir-42	II	11889765-11889784		11889711-11889790	8.85
mir-43	II	11889864-11889886		11889816-11889895	9.80
mir-44	II	11889977-11889997		11889926-11890005	7.04
mir-46	III	11994834-11994855		11994782-11994861	5.66
mir-47	X	13674086-13674107		13674036-13674115	7.94
mir-48	V	14209762-14209784	(-)	14209717-14209796	8.69
mir-49	X	9754719-9754740		9754666-9754745	6.30
mir-50	I	98223-98246 in Y71G12B		98215-98295	5.00
mir-51	IV	9361617-9361639	(-)	9361573-9361652	4.33
mir-53	IV	9363196-9363219	(-)	9363153-9363232	4.41
mir-54	X	12897785-12897808	(-)	12897781-12897860	7.32
mir-55	X	12897616-12897638	(-)	12897606-12897685	8.91
mir-56	X	12897480-12897501	(-)	12897476-12897555	6.30
mir-56b	X	12897522-12897544	(-)	12897476-12897555	6.30
mir-57	II	7850475-7850498	(-)	7850431-7850510	3.96
mir-58	I	16248-16269 in Y67D8A		16198-16282	7.15
mir-59	IV	9728930-9728952	(-)	9728922-9729001	4.85
mir-60	II	6328662-6328684	(-)	6328728-6328659	8.40
mir-61	V	11628209-11628229	(-)	11628197-11628276	8.08
mir-62	X	12445416-12445437		12445376-12445455	4.48
mir-64	III	93001-93023 in Y48G9A		92993-93072	4.61
mir-65	III	93151-93173 in Y48G9A		93140-93219	4.43
mir-66	III	93256-93278 in Y48G9A		93248-93327	3.77
mir-67	III	4744361-4744384	(-)	4744354-4744433	5.27
mir-70	V	6538459-6538481	(-)	6538448-6538527	4.80
mir-71	I	7704477-7704495	(-)	7704505-7704436	5.74
mir-72	II	2452852-2452871		2452841-2452920	5.12
mir-73	X	2105074-2105096		2105026-2105105	4.95
mir-74	X	2105351-2105372		2105301-2105380	6.57
mir-75	X	2108762-2108783		2108716-2108795	6.04
mir-76	III	2006970-2006991		2006915-2006994	7.20
mir-77	II	12519222-12519243		12519171-12519250	7.45
mir-79	I	7657496-7657517		7657482-7657561	4.25
mir-80	III	7685424-7685446	(-)	7685497-7685418	5.81
mir-81	X	2167389-2167410		2167336-2167415	5.73
mir-82	X	2171524-2171545	(-)	2171519-2171588	5.11
mir-83	IV	6202645-6202666		6202596-6202675	5.41
mir-84	X	15895740-15895761	(-)	15895695-15895764	4.59
mir-85	II	8393532-8393555		8393486-8393565	6.47
mir-86	III	1842256-1842278	(-)	1842251-1842320	3.91
mir124	IV	10276532-10276552		10276524-10276593	5.83
mir228	IV	4250414-4250436		4250407-4250486	4.68
mir230	X	5538452-5538474		5538401-5538480	6.96
mir231	III	6362417-6362440	(-)	6362411-6362490	8.09
mir232	IV	9329727-9329749	(-)	9329723-9329802	5.77
mir233	X	11844227-11844249	(-)	11844222-11844291	3.35
mir234	II	14461938-14461958	(-)	14461921-14462000	7.99
mir235	I	4504454-4504475	(-)	4504452-4504521	4.25
mir236	II	7030136-7030158	(-)	7030116-7030195	7.87

Table 1. (continued)

Gene	Chromosome and Location			WFS	Zscr
mir237	X	7902163-7902186		7902151-7902230	8.96
mir238	III	7687482-7687504	(-)	7687478-7687547	3.10
mir239a	X	11559966-11559988		11559961-11560040	4.79
mir239b	X	11558868-11558890	(-)	11558821-11558900	6.70
mir243	IV	3199781-3199803		3199731-3199810	12.62
mir244	I	3011309-3011332	(-)	3011261-3011340	4.51
mir245	I	6233176-6233197		6233133-6233202	5.78
mir246	IV	9312591-9312612		9312543-9312622	4.50
mir247	X	4509401-4509423		4509351-4509430	7.05
mir248	X	2002036-2002058	(-)	2002031-2002100	5.53
mir249	X	2745290-2745311	(-)	2745281-2745360	8.58
mir250	V	11628066-11628087	(-)	11628057-11628136	8.42
mir251	X	10746280-10746303		10746276-10746345	5.28
mir252	II	11446827-11446849	(-)	11446766-11446845	4.04
mir253	V	5606449-5606469	(-)	5606408-5606487	5.89
mir254	X	11075086-11075108	(-)	11075081-11075160	11.00

For example, miRNA mir-46 is encoded in the region $11994834 - 11994855$ of chromosome III and mir-48 is encoded in the reverse complementary sequence (RCS) $14209762 - 14209784$ of chromosome V of *C.elegans*. Figure 1 graphically depicts the observed distributions of the scores $Zscr_e$ computed in the two genomic sequences of 2500-nt that contain *C.elegans* mir-46 and mir-48 genes, respectively. For the segment $11993701 - 11996200$ of chromosome III, the maximal $Zscr_e$ was 5.66 and found in the WFS $11994782 - 11994861$ of chromosome III. The 22-nt mir-46 was located at the right arm of the distinct stem-loop folded by WFS $11994782 - 11994861$. Among them, 19 nt out of 22-nt were in the base-pairs. Similarly, the maximal $Zscr_e$ was 8.69 and found in the WFS $14209717 - 14209796$ in the segment $14208531 - 14211030$ of chromosome V. The RCS of mir-48 was situated in the right arm of the folded stem-loop and 20-nt out of 23-nt mir-48 were in the base-pairing region of the distinct WFS $14209717 - 14209796$.

What is the general behavior of E_{diff} or $Zscr_e$ in a random sample that is associate with the real biological sequence? To estimate the uncertainty of E_{diff} in a random sample we performed an extensive statistical simulations for 50 randomly shuffled sequences of the segment $11993701 - 11996200$ of chromosome III, and the segment $14208531 - 14211030$ of chromosome V, respectively. The quantitative measures E_{diff} and $Zscr_e$ were computed by using same parameters as used in *C.elegans*. In each random test, the total length of random sequences were 125000 nt and we had 24250 observations of $Zscr_e$. The distribution of $Zscr_e$ computed in random sequences are showed in Figure 2. In the statistical simulation of the segment $11993701 - 11996200$ of chromosome III, $Zscr_e$ scores ranged from -1.50 to 5.57. There were 105, 17 and 2 observations whose $Zscr_e$ values were equal to or greater than 3.5, 4.5 and 5.5, respectively. For the segment $14208531 - 14211030$ of chromosome V, $Zscr_e$ scores ranged from -1.52 to 6.23 in the random test. There were 102, 17 and 1 observations whose $Zscr_e$ values were greater than 3.5, 4.5 and

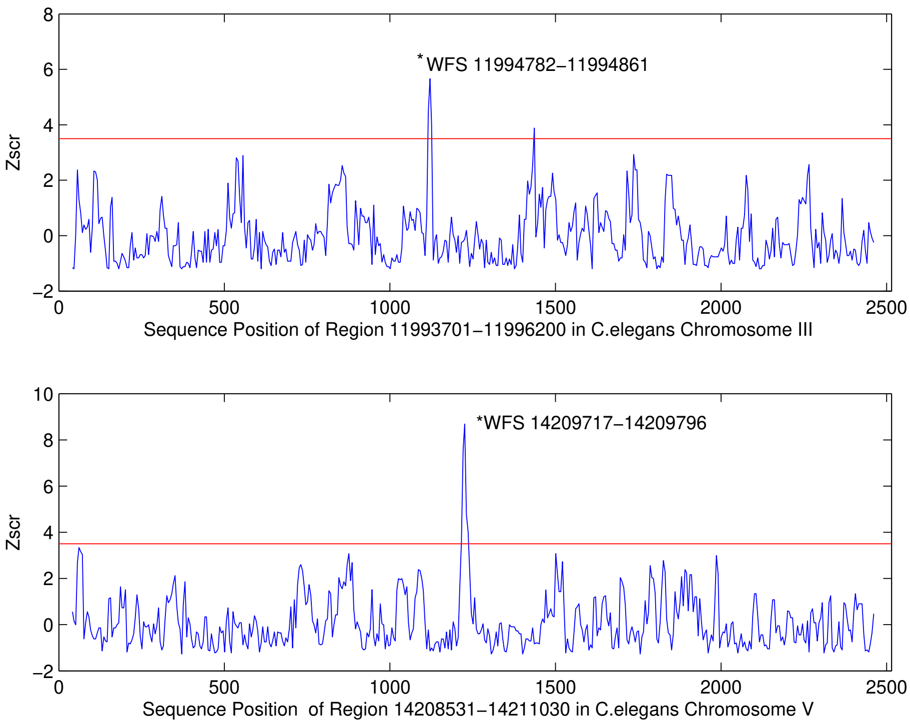

Fig. 1. Distributions of $Zscr_e$ scores computed in the two genomic sequences of segment 11993701-11996200 of chromosome III (top) and segment 14208531-14211030 of chromosome V (bottom) of *C.elegans*. Each plot was made by plotting $Zscr_e$ against the position of the middle base in the window of 80 nt. The detected WFS including mir-46 (top) and mir-48 are asterisked in the plot.

5.5, respectively. On average, we can expect to have two observations whose $Zscr_e \geq 3.5$ in a 2500-nt random sequence, and have 1.5 observations whose $Zscr_e \geq 5.5$ in a random sequence of 100000-nt. From the statistical simulations we can set a reasonable $Zscr_e^0 = 5.5$ for predicting the statistically significant WFS in *C. eleg.* The statistical analysis of random samples also indicate that the most of WFS elements listed in Table 1 are statistically very significant. Those detected WFS elements associated with miRNAs represent a well-ordered structural feature that can not be expected by chance.

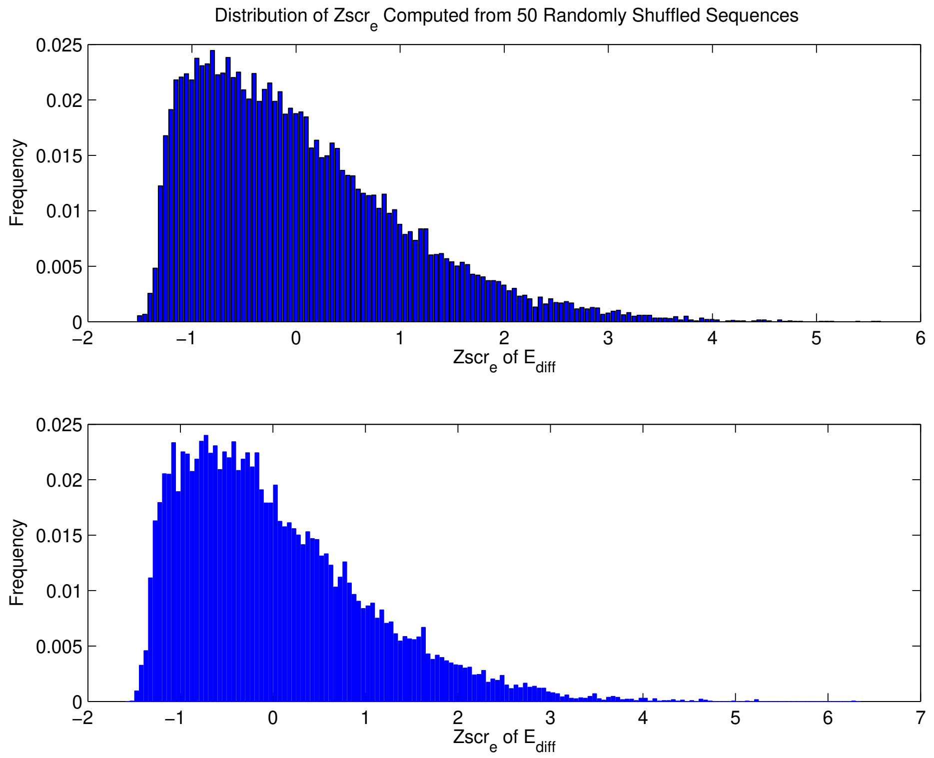

Fig. 2. Empirical probability density functions of $Zscr_e$ scores computed from 50 random sequences. The $Zscr_e$ scores were computed from the randomly shuffled sequences of the segment 11993701-11996200 of *C.elegans* chromosome III are displayed in the top and those computed from the segment 14208531-14211030 of *C.elegans* chromosome V are depicted in the bottom. The empirical bar functions are plotted with step size of $Zscr_e = 0.05$. $Zscr_e$ were computed by the same parameters as used in real biological sequence.

3.2. *Structural features of the miRNA Precursor*

It is know that ∼80-nt miRNA precursors often form a distinct fold-back structure in which most of the 21−24 nt miRNAs are in the double helical stem. Moreover, the fold-back stem-loop structure is also highly conserved across the divergent species. Figure 3 shows the conserved stem-loop structure of *let-7* miRNA precursors found in *C.elegans, D.melanogaster* and human. The high structure conservation can be measured by MSS scores. The computed MSS between the two *let-7* RNA structures of *C.elegans* and *D.melanogaster* was 247. Similarly, we had $MSS = 228$, and 212 computed by the structure comparison between the *let-7* precursors of *C.elegans* and

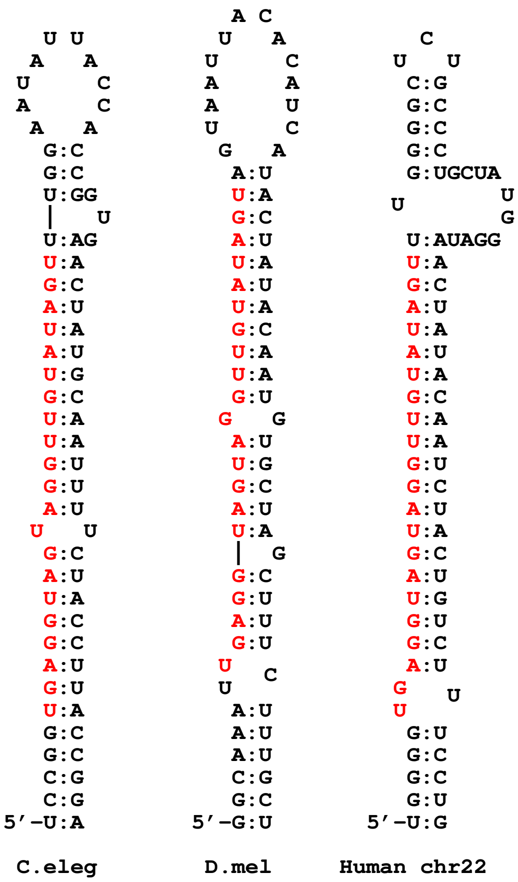

Fig. 3. Fold-back stem-loop structures of *C.elegans*, *D.melanogaster* and human *let-7* RNA precursors. The 21-nt *let-7* RNA is shown by red (lighter) color.

human, and *D.melanogaster* and human, respectively. Based on the basic information, the prediction of the conserved WFS related to the *let-7* gene in the genomic sequence of *C.elegans* was set by the two conditions, $Zscr_e \geq 5.5$ and $MSS \geq 228$.

Using the window of 80-nt, we identified 1853, 1625, 1278, 1517, 1708 and 1256 WFS elements in the chromosome I, II, III, IV, V and chromosome X, respectively by the threshold, $Zscr_e^0 = 5.5$. Using both thresholds of $Zscr_e$ and MSS, we detected 8, 8, 9, 10, 5, and 4 WFS elements in chromosome I, II, III, IV, V and X, respectively. The detected WFS have similar well-ordered conformation as that found in *let-7* precursors (see Table 2). Table 2 gives the conserved WFS elements from chromosome IV by our computation.

Table 2. Distinct WFS that has conserved structural feature of let7 precursor computed by EDscan and rna_match from Chromosome IV with a window size of 80-nt.

```
-------------------------------------------------------------------------------------
Folding region: 14392174~14392258; Zscr = 5.83, MSS = 244.0
aaatttTCTTAGGAAAAGTGTTAATTGAAAAGTTtTAGATaaagttataATCTAcAACTTTTTAGTTGAtCACTTTTTTgTGAGA
       1   14392180   14392258      5
       2   14392185   14392252      9
       3   14392194   14392242     14
       4   14392209   14392227      5
Folding region: 15327884~15327968; Zscr = 8.08, MSS = 243.0
CGGCCGTCAGTTTCCGAGTTTaGCcACTCATTaTACaatgtattattaaGTAcAGTGAGTaGCcAAACTTGGAAATTGACGGCCG
       1   15327884   15327968     21
       2   15327906   15327946      2
       3   15327909   15327943      7
       4   15327917   15327935      3
Folding region: 15247729~15247813; Zscr = 8.08, MSS = 243.0
CGGCCGTCAGTTTCCGAGTTTaGCcACTCATTaTACaatgtattattaaGTAcAGTGAGTaGCcAAACTTGGAAATTGACGGCCG
       1   15247729   15247813     21
       2   15247751   15247791      2
       3   15247754   15247788      7
       4   15247762   15247780      3
Folding region:  2660337~ 2660426; Zscr = 8.88, MSS = 239.0
AGTGTCGGATGGGagCcAAGTTTGCACTAAATAGTGAcataccctaTCGCaaTATTTAGTGCAAACTTtGtcCCCGTCCGACACTttttg
       1    2660337    2660421     13
       2    2660352    2660406      1
       3    2660354    2660404     16
       4    2660370    2660386      4
Folding region: 14647174~14647258; Zscr = 8.02, MSS = 237.0
ggtTCCGCGCGCCGCtATGTTTAACTCGCagcGGGCGgagacagcttgccaCGCCCacaGCGAGTTAAACATcGCGGCGCGCGGA
       1   14647177   14647258     12
       2   14647190   14647245     13
       3   14647206   14647229      5
Folding region:  3700280~ 3700364; Zscr = 8.86, MSS = 234.0
ATTGTTCGAAAGTTGAaAATTacCGGTGAAAATTGCCaaaaaattgacGGCAATTTTCATCGtcAATTcTCAACTTTCGGACAGT
       1    3700280    3700364     16
       2    3700297    3700347      4
       3    3700303    3700341     14
Folding region: 12303329~12303408; Zscr = 7.79, MSS = 231.0
TCAAGTAAtGTAGcGAaATGTATTTAAATACATTTGTGacgtCACAAATGTATTTAAATACATgTTtTTATTTACTTGAa
       1   12303329   12303407      8
       2   12303338   12303399      4
       3   12303343   12303394      2
       4   12303346   12303391     21
Folding region: 15150209~15150293; Zscr = 6.77, MSS = 229.0
gTTCCGCGCGCCGCTaTGTTTAACTCGCagcaGGCGgagacagcttgccaCGCCcacaGTGAGTTAAACGcAGCGGCGCGCGGAA
       1   15150210   15150293     14
       2   15150225   15150278     12
       3   15150241   15150262      4
Folding region: 13596834~13596913; Zscr = 10.71, MSS = 229.0
GCGTGAAACTGTAATTTTcTGTaCCAAAAAatTCAAaaaccctgcaTTGAcTTTTTGGACAaAAAATTACAGTTTCACGC
       1   13596834   13596913     18
       2   13596853   13596894      3
       3   13596857   13596891      7
       4   13596866   13596883      4
Folding region: 11328499~11328583; Zscr = 8.19, MSS = 229.0
TTTTtAGGTTAATTAAaCATTATATCATCAAAAAacGAaaaatttgTCaggTTTTTGATGATATAATGgTTAATTGACTTcAAGA
       1   11328499   11328583      4
       2   11328504   11328578     11
       3   11328516   11328566     17
       4   11328535   11328546      2
-------------------------------------------------------------------------------------
```

In Table 2, we first list the WFS sequence in which the capital letters represent the base pairing regions. The WFS sequence is followed by a simple region table where the 5' and 3' positions of the stems in the sequence are listed in the columns 2 and 3 and the size of the base pairing in the stem is listed in the column 4. Figure 4 depicts partial fold-back stem-loops listed in Table 2. It is clear that the detected structures are phylogenetically conserved and uniquely folded. Statistical inference for the well-ordered structures indicates that these are not expected in chance.

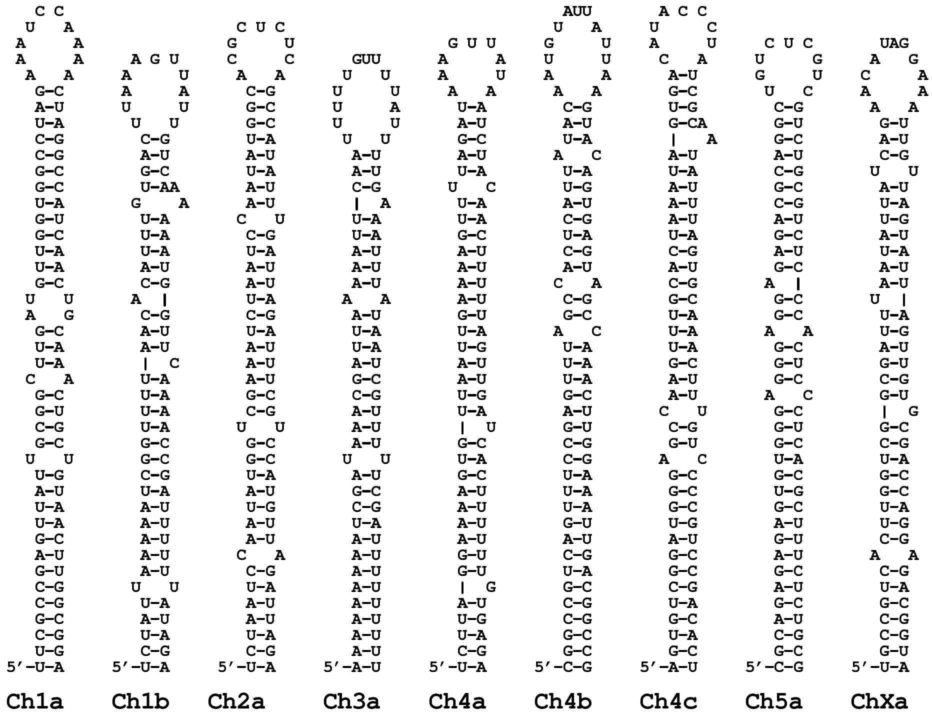

Fig. 4. Examples of conserved WFS structures detected in *C.elegans*. They are partial data listed in Table 2. The distinct stem-loop structures have the conserved structure feature found in the *let-7* precursor. Stem-loop Ch1a is folded by WFS 10582568-10582652 of chromosome I whose $Zscr_e$ is 6.74 and MSS is 262. Stem-loop Ch1b is folded by WFS 12684236-12684315 in chromosome I whose $Zscr_e$ is 5.99 and MSS is 239. Stem-loop Ch2a is folded by WFS 14797341-14797425 in chromosome II whose $Zscr_e$ is 6.59 and MSS is 250. Stem-loop Ch3a is folded by WFS 3398800-3398879 of chromosome III whose $Zscr_e$ is 6.17 and MSS is 240. Stem-loop Ch4a, Ch4b and Ch4c are folded by WFS 14392174-14392258, 15327884-15327968 and 2660337-2660426 of chromosome IV, respectively. Their $Zscr_e$ scores are 5.83, 8.08, and 8.88, and MSS values are 244, 243 and 239, respectively. Stem-loop Ch5a is folded by WFS 17495056-17495135 in chromosome V whose $Zscr_e$ is 8.0 and MSS is 235. Stem-loop ChXa is folded by WFS 5538401-5538480 of chromosome X whose $Zscr_e$ is 6.96 and MSS is 253.

The MSS threshold used here is an empirical one and it should be justified case by case. In general one should choose a score of about 10-15% lower than the self similarity score of the given miRNAs. Lowering this score will produce more but less conserved WFS elements. In addition, those WFS with high Zscr that are not conserved are only not closely related to the given miRNAs. Their potential biological functions can not be ruled out.

Similarly we also found other distinct well-ordered structures that were related to the other known miRNAs listed in Table 1 by the same procedure as used above. Those detected, statistically significant WFS elements may be potential candidates of undiscovered miRNAs or other functional elements. The number of known miRNAs is expanding rapidly. It is a great challenge to discover those miRNAs by computational methods in the post-genomic era. Although some computational approaches for predicting functional RNAs in genomic sequences have been proposed [24,25], we still need more sophisticated computational tools. This study showed that most of known *C.elegans* miRNAs were associated with those statistically significant WFS. Our method can predict *let-7* and other experimentally determined miRNAs. Given the distinct morphology of the fold-back structure of a specific miRNA, we can search for its homologue by performing the structural comparison between the specific miRNA structure and the structures of those potential WFS segments detected in *C.elegans* genome and other genomic sequences. Once those potential homologous RNAs are determined, we can use sequence search to find those with conserved subsequences. Currently, predictions of those potential functional structures in *C.elegans* and other genomic sequences are being conducted on a large scale in our laboratory.

4. Conclusion

The miRNAs in size from 21 to 25 nt have been predominant in eukaryotes. In this study, we proposed a general approach with the combination of EDscan (http://protein3d.ncifcrf.gov/ shuyun/edscan.html) and rna_match (http://www.csd.uwo.ca/faculty/kzhang/rna/rna_match.html) to search for conserved fold-back structures of miRNA precursors. Our statistical simulation indicates that such fold-back stem-loops are statistically significant and they are not expected by chance. The distinct structure is both thermodynamically stable and uniquely folded. Using the approach, we discovered a number of the potential fold-back structures in *C.elegans* genome. Our method will help to find miRNAs and other interesting structural features hidden in the enormous volume of the complete genome.

Acknowledgments

We thank anonymous reviewers for their suggestions. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. K.Z. is supported in part by Natural Sciences and Engineering Research Council of Canada grant OGP0046373, a research fellowship from Simon Fraser University and a Sharcnet research fellow-

ship.

References

1. J.S. Mattick, Non-coding RNAs: the architects of eukaryotic complexity, *EMBO Rep.* **2** (2001) 986-991.
2. G. Storz, An expanding universe of noncoding RNAs, *Science* **296** (2002) 1260-1263.
3. R.H. Plasterk, RNA silencing: the genome's immune system, *Science* **296** (2002) 1263-1265.
4. P.D. Zamore, Ancient pathways programmed by small RNAs, *Science* **296** (2002) 1265-1269.
5. R.F. Gesteland and J.F. Atkins, *The RNA World*, (Cold Spring Harbor Lab. Press, New York, 1993).
6. M.W. Hentze, S.W. Caughman, J.L. Casey, D.M. Koeller, T.A. Rouault, J.B. Harford, and R.D. Klausner, A model for structure and functions of iron-responsive elements, *Gene* **72** (1988) 201-208.
7. A.E. Pasquinelli, *et al.*, Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA, *Nature* **408** (2000) 86-89.
8. M.H. Malim, J. Hauber, S.-Y. Le, J.V. Maizel Jr., and B.R. Cullen, The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA, *Nature* **338** (1989) 254-257.
9. P.M. Macdonald and C.A. Smibert, Translational regulation of maternal mRNAs, *Curr. Opin. Genet. Dev.* **6** (1996) 403-407.
10. C.U.T. Hellen and P. Sarnow, Internal ribosome entry sites in eukaryotic mRNA molecules, *Genes & Development* **15** (2001) 1593-1612.
11. I. Brierley, P. Digard, and S.C. Inglis, Characterization of an efficient coronavirus ribosomal frameshifting signal requirement for an RNA pseudoknot, *Cell* **57** (1989) 537-547.
12. D.E. Draper, Strategies for RNA folding, *Trends Biochem. Sci.* **21** (1996) 145-149.
13. S.-Y. Le, K. Zhang, and J.V. Maizel Jr., RNA molecules with structure dependent functions are uniquely folded, *Nucleic Acids Res.* **30** (2002) 3574-3582.
14. V. Ambros, *et al.*, A uniform system for microRNA annotation, *RNA* **9** (2003) 277-279.
15. L.P. Lim, N.C. Lau, E.G. Weinstein, A. Abdelhakim, S. Yekta, M.W. Rhoades, C.B. Burge, and D.P. Bartel, The microRNAs of *Caenorhabditis elegans*, *Genes & Development* **17** (2003) 991-1008.
16. J.C. Venter, *et al.*, The sequence of the human genome, *Science* **291** (2001) 1304-1351.
17. S.Y. Le, J.H. Chen, D. Konings, and J.V. Maizel Jr., Discovering well-ordered folding patterns in nucleotide sequences, *Bioinformatics* **19** (2003) 354-461.
18. S.Y. Le, J.H. Chen, D. Konings, and J.V. Maizel Jr., Local well-ordered folding sequences in Ferritin and transferrin receptor mRNAs, *Proceedings of the international conference on METMBS'02* (2002) 41-47.
19. G.D. Collins, S.Y. Le, and K. Zhang, A new algorithm for computing similarity between RNA structures, *Information Sciences* **139** (2001) 59-77.
20. M. Zuker and P. Steigler, Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information, *Nucleic Acids Res.* **9** (1981) 133-149.
21. D.H. Mathews, J. Sabina, M. Zuker and D.H. Turner, Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure, *J. Mol. Biol.* **288** (1999) 911-940.
22. K. Zhang, Computing similarity between RNA secondary structures, *Proceedings of IEEE International Joint Symposia on Intelligence and Systems* (1988) 126-132.

23. K. Zhang, L. Wang and B. Ma, Computing similarity between RNA structures, *Proceedings of the Tenth Symposium on Combinatorial Pattern Matching. Springer-Verlag's Lecture Notes in Computer Science 1645* (1999) 281-293.
24. E. Rivas and S.R. Eddy, Noncoding RNA gene detection using comparative sequence analysis, *BMC Bioinformatics* **2** (2001) 8-26.
25. R.J. Carter, I. Dubchak and S.R. Holbrook, A computational approach to identify genes for functional RNAs in genomic sequences, *Nucleic Acids Res.* **29** (2001) 3928-3938.