BIOINFORMATICS
Vol. 19 no. 3 2003, pages 354–361
DOI: 10.1093/bioinformatics/btf826

Discovering well-ordered folding patterns in nucleotide sequences

Shu-Yun Le[1,*], Jih-H. Chen[2], Danielle Konings[1] and Jacob V. Maizel, Jr.[1]

[1]Laboratory of Experimental and Computational Biology, NCI Center for Cancer Research, National Cancer Institute, NIH, Bldg 469, Room 151, Frederick, MD 21702 and [2]Advanced Biomedical Computing Center, SAIC-NCI/FCRDC, P.O. Box B, Frederick, MD 21702, USA

Received on May 6, 2002; revised on July 23, 2002; accepted on July 26, 2002

ABSTRACT

Motivation: Growing evidence demonstrates that local well-ordered structures are closely correlated with *cis*-acting elements in the post-transcriptional regulation of gene expression. The prediction of a well-ordered folding sequence (WFS) in genomic sequences is very helpful in the determination of local RNA elements with structure-dependent functions in mRNAs.

Results: In this study, the quality of local WFS is assessed by the energy difference (E_{diff}) between the free energies of the global minimal structure folded in the segment and its corresponding optimal restrained structure (ORS). The ORS is an optimal structure under the condition in which none of the base-pairs in the global minimal structure is allowed to form. Those WFSs in HIV-1 mRNA, various ferritin mRNAs and genomic sequences containing *let-7* RNA gene were searched by a novel method, ed_scan. Our results indicate that the detected WFSs are coincident with known Rev response element in HIV-1 mRNA, iron-responsive elements in ferritin mRNAs and small *let*-7 RNAs in *Caenorhabditis elegans*, *Caenorhabditis briggsae* and *Drosophila melanogaster* genomic sequences. Statistical significance of the WFS is addressed by a quantitative measure $Zscr_e$ that is a z-score of E_{diff} and extensive random simulations. We suggest that WFSs with high statistical significance have structural roles involving their sequence information.

Availability: The source code of ed_scan is available via anonymous ftp as ftp://ftp.ncifcrf.gov/pub/users/shuyun/scan/ed_scan.tar.

Contact: shuyun@orleans.ncifcrf.gov

INTRODUCTION

RNA and DNA are conformationally polymorphic molecules. Numerous experimental results have shown that RNAs perform a wide range of functions in biological systems. Though single-stranded regions exist in most RNAs, local well-ordered structures in RNA sequences correlate closely with functions such as the control of replication, mRNA processing, metabolism and translation (Simons and Grunberg-Manago, 1998; Gray and Wickens, 1998; Bashirullah *et al.*, 1998). It has become clear that the interactions between RNA and RNA, as well as RNA and protein play a crucial role in the regulatory control. Knowledge of the conformation of local RNA structures is essential for our understanding of the regulatory mechanisms. Therefore functional prediction for distinct folding patterns in sequences is an important goal of genomic sequence analysis. Computational searches for potential, functional structured elements in genomic sequences are highly desirable in the post-genomic era.

Previous studies (Le *et al.*, 1990a; Dayton *et al.*, 1992; Forsdyke, 1995; Patzel and Sczakiel, 1997) suggested that structured, functional RNA elements were often thermodynamically more stable than those that would be anticipated for equivalent random sequences. In the study of RNA folding energy landscapes, Chen and Dill (2000) indicated that RNA folding energy landscapes might be quite bumpy and rugged. The RNA folding often involves complex intermediate states and may have many low energy states. It is also true that some functional RNA molecules, such as tRNA and RNase P RNA are often not represented by the most thermodynamically stable RNA folding. Recently, Rivas and Eddy (2000) suggested that the stability of some non-coding RNA secondary structures was not sufficiently different from the predicted stability of random sequences. It has been shown that some distinct loop sequences and specific combinations of base pairings in stem-loops are functional RNA structural motifs (Hermann and Patel, 1999). Thus, functional structural RNA molecules possess well-ordered conformations that are both thermodynamically stable and

*To whom correspondence should be addressed.

Published by Oxford University Press

uniquely folded (Draper, 1996; Schultes *et al.*, 1999).

In this study, we search for these well-ordered folding sequences (WFS) by a newly developed method, ed_scan. We use a quantitative measure, E_{diff}, to characterize the thermodynamic stability and well-ordered conformation of a local RNA secondary structure. E_{diff} is the difference of free energies between the global minimal energy structure and its corresponding optimal restrained structure (ORS). We further introduce a standard z-score, $Zscr_e$ to analyze E_{diff} of local segments in a long sequence. Computed WFSs in HIV-1 and ferritin mRNAs with $Zscr_e$ significantly different from the bulk distribution correlate with previously known Rev responsive elements (RRE) in the HIV-1 (Malim *et al.*, 1989) and iron-responsive elements (IRE) in human and other species (Hentze *et al.*, 1988). Using ed_scan, we also find *let-7* RNA genes (Pasquinelli *et al.*, 2000) in the sequences of *Caenorhabditis elegans* (*C.elegans*), *Caenorhabditis briggsae* (*C.briggsae*), and *Drosophila melanogaster* (*D.melanogaster*). Statistical significance of the detected WFS is assessed by random simulations. Our results show that statistical extremes detected in the wild-type sequences correlate with functional RNA elements and are not expected to occur by chance. The method is generally suitable to search for potential, structured functional elements in a genomic sequence.

SYSTEMS AND METHODS

In RNA folding, energies involved in the formation of secondary structure are greater than those involved in tertiary interactions (Tinoco and Bustamante, 1999). The energetic contributions of the secondary and tertiary structural elements in RNA are also separable. In this study, we compute the lowest free energy of folded local RNA segments by a dynamic programming algorithm (Zuker, 1989) and Turner energy rules (Mathews *et al.*, 1999). The measure E_{diff} of a segment is defined as:

$$E_{\mathrm{diff}} = E_f - E$$

where E is the lowest free energy of the local folded segment, E_f is also the lowest free energy of the ORS that is computed under the condition in which all base-pairs formed in the global minimal structure are prohibited. Consequently, the measure E_{diff} signifies the stability and uniqueness of the predicted RNA secondary structure from the local segment. The greater the E_{diff} of the segment is, the more well-ordered the folded RNA structure is expected to be. To further analyze E_{diff} we introduce a standardized measure $Zscr_e$

$$Zscr_e = (E_{\mathrm{diff}} - E_{\mathrm{diff}}(w))/std_w$$

where $E_{\mathrm{diff}}(w)$ and std_w are the mean and standard deviation, respectively, of the E_{diff} scores computed by sliding a fixed-length window in steps of a few nucleotides (nt) from 5′ to 3′ along the sequence.

In searching for distinct WFSs in a sequence the following steps are generally employed. (i) $Zscr_e$ values are computed by sliding a window with a proper size, for instance 100 nt, along the sequence. The potential interesting regions with high $Zscr_e$ are chosen based on the profile of $Zscr_e$ in the sequence. (ii) The precise locations of those potential targets in which the folded structure is highly well-ordered are inferred by an extended search in the regions determined from the step 1. In the extended search, the window size is systematically changed with a range of sizes (e.g. 60–300) and the maxima of $Zscr_e$ are extracted to determine the optimized WFS. If the method is applied to single-stranded DNA sequences, the folding energy parameters derived from experimental thermodynamic data of DNA (SantaLucia, 1998) are used to compute the lowest free energy of a folded DNA fragment.

The computer program ed_scan is designed to calculate the score $Zscr_e$ of a local segment in a sequence. The program is based on the dynamic programming algorithm (Zuker, 1989; Mathews *et al.*, 1999) and implemented in Fortran 90 running on Unix. The program has been tested and used on a Compaq/DEC Alpha 8400/625 EV56 platform, SGI/Octane and SGI/Onyx platform with IRIX 6.5. All statistical analyses for our data were performed in this study using the Statistics Toolbox of MATLAB software package (http://www.mathworks.com).

RESULTS AND DISCUSSION

WFS patterns are closely associated with the reported functional RNA elements in human immunodeficiency virus

The outline landscape of the distribution for WFSs in the mRNA sequence of HIV-1 isolate BRU (Accession number: K02013) was made by computing $Zscr_e$ using a window of 100 nt. In the computation, the window was first set at the 5′ end of the sequence and then was moved by 5 nt in each time until it reached to the 3′ end of the sequence. For the $Zscr_e$ data only five domains (labeled by peaks A–E) whose $Zscr_e$ values are equal to or greater than 3.5 (Figure 1). The domains A and E are scattered in the both 5′ and 3′ non-coding regions, and the domain B is located at the gag-pol overlapping region. The domain C is located at the envelope coding region just 50 nt downstream from the cleavage site of the outer membrane protein (OMP) and transmembrane protein (TMP). Domain D, centered at the position 8746, is located at the coding region of nef protein. Except the peak D, the other four peaks are closely correlated with known functional RNA elements in HIV-1, such as *trans*-activation response (TAR) and RRE and *cis*-acting element of the regulation of gag-pol frameshifting.

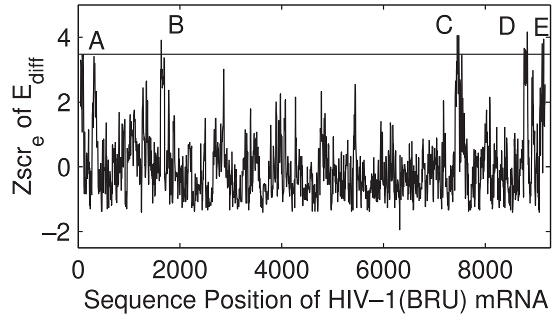

Fig. 1. $Zscr_e$ of the energy difference of the lowest free energies between the optimized and corresponding optimal restrained structure of a local segment in the complete HIV-1 (isolate BRU) mRNA sequence. The plot was produced by plotting $Zscr_e$ of a 100 nt segment against the position of the middle base in the segment as it was moved successively by five nt from 5′ to 3′ along the sequence. Well-ordered folding regions are labeled by letters A–E.

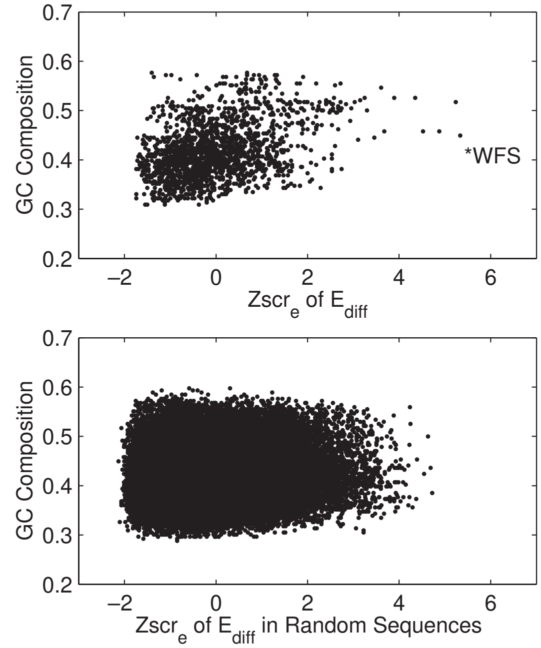

Fig. 2. Relationships between GC composition and $Zscr_e$ of local segments of 236 nt computed in wild type sequence of HIV-1 mRNA (top) and 30 randomly shuffled sequences (bottom) of the wild type sequence. $Zscr_e$ values were computed by moving a 236-nt window in steps of 5 nt from 5′ to 3′ along the sequences. The significant WFS 7361–7596 ($Zscr_e = 5.33$, the maximal score as asterisked in the plot) is apparently separated from the bulk distribution in the wild type sequence.

The optimized boundaries of these regions are further delimited by an extensive search. For the region corresponding to the RRE element we set the search domain from position 6501 to position 8500 centered on the position of peak C. This domain includes $\sim$850 nt upstream and $\sim$1150 nt downstream from the cleavage site of the OMP/TMP for the HIV-1 (BRU) mRNA. The size of the scanning window is systematically changed from 80 to 350 nt by a step of 10 nt, along with 370, 390 and 236 nt. For each fixed-length window, the score $Zscr_e$ was computed by sliding the window in steps of 5 nt from 5′ to 3′ along the sequence. Our results indicate that there are only 5 maximal $Zscr_e$ ($\geqslant 5.0$) that are detected by five windows of 160, 170, 236, 240 and 340 nt, respectively (detailed data see ftp://ftp.ncifcrf.gov/pub/users/shuyun/scan/Supp/hivbru_w80-390.ps). They correspond to segments 7391–7550 ($Zscr_e = 5.11$), 7391–7560 ($Zscr_e = 5.38$), 7361–7596 ($Zscr_e = 5.22$), 7361–7600 ($Zscr_e = 5.08$) and 7311–7650 ($Zscr_e = 5.01$). It is clear that all five WFSs are overlapped and coincide with the known HIV-1 RRE element (Malim *et al.*, 1989; Le *et al.*, 1990a; Mann *et al.*, 1994). The WFS of region 7361–7596 detected by the window of 236 nt is completely coincident with the RRE determined by mutagenesis experiments (Malim *et al.*, 1989).

Statistics of $Zscr_e$ in human immunodeficiency virus

Statistical analysis indicates that the $Zscr_e$ data show asymmetry with sample mean, $m = 0$, sample standard deviation, $std = 1.0$. The distribution of $Zscr_e$ is skewed toward the positive direction and is not well fitted by a normal distribution. For example, the sample coefficients of skewness (k) is 0.985 for the data $Zscr_e$ computed by sliding the window of 236 nt along the complete HIV-1 (BRU) mRNA. The observed distribution of $Zscr_e$ ranged from -1.74 to 5.33, and the distribution of GC composition ranged from 0.31 to 0.58. It is evident that there is no $Zscr_e$ that is apparently separated from the bulk distribution in the negative direction. However, the WFS 7361–7596 ($Zscr_e = 5.33$) is clearly separated from the bulk distribution in the positive direction, and is statistically significant (see the top of Figure 2).

In order to evaluate the statistical significance of the distinct WFS we performed an extensive computational experiment for 30 randomly shuffled sequences of HIV-1 (BRU) mRNA. It was important that randomizations were done by shuffling so that the same base composition and length as the wild type sequence were maintained. The $Zscr_e$ values of local segments were computed by sliding a window of 236 nt along the 30 random sequences using the same parameters as mentioned above. In random tests, the total length of random sequences were 276 870 nt and we had 53 970 observations of $Zscr_e$. The mean and standard deviation for each sample were zero and 1.0, respectively. $Zscr_e$ scores ranged from -2.13 to 4.72. There were 103 or 22 observations whose $Zscr_e$ values were greater than 3.5 or 4.0 in 53 970 observations. Only 4 out of total 53 970 observations of $Zscr_e$ were greater than 4.50. The observed probabilities of WFS with $Zscr_e \geqslant$ 4.0, 3.5, 3.0 and 2.5 are less than 0.0004, 0.0019, 0.0066 and 0.0172 in the random test, respectively. From the random test we show that the WFS 7361–7596 detected in HIV-1(BRU) mRNA is statistically significant and can not be expected by chance, reinforcing the observation that high $Zscr_e$ represents a well-ordered conformation.

Relationships between GC compositions and $Zscr_e$ of the local segments computed in the HIV-1(BRU) mRNA and 30 corresponding random sequences were summarized and shown in Figure 2. The significant WFS can be apparently separated from the bulk distribution in the positive direction and is not sensitive to the GC composition of the local segment. Fragments with different GC compositions ranging from 0.32 to 0.55 have similar $Zscr_e$, providing evidence that $Zscr_e$ values are not sensitively affected by GC compositions.

WFSs are coincident with the reported IRE found in ferritin mRNAs

Ferritin biosynthesis is regulated translationally by a conserved translational regulatory sequence in the mRNA 5′- untranslated region (5′-UTR) called the IRE. The IRE has been demonstrated to function by forming a distinct stem-loop structure including a conserved bulge and hairpin loop (Hentze *et al.*, 1988). RNA folding studies of the IREs have shown that the computed secondary structures are only moderately stable in general (Hentze *et al.*, 1988). Using the previously developed methods SEGFOLD and SIGSTB (Le *et al.*, 1990b) we observed that the computed significance scores of most of these IRE segments of 24 ferritin mRNAs from human and other species were not sufficiently more stable than those from corresponding random sequences. Thus, the known IREs can not be reliably detected in most of 24 ferritin mRNAs based on the quantitative measure of thermodynamic stability only (Table 1).

With the new measure $Zscr_e$, we searched for IRE of the 24 ferritin mRNAs by ed_scan. In the computation the size of the scanning window was systematically changed from 25 to 80 nt by increasing 5 nt each time. $Zscr_e$ values were computed by sliding these windows in steps of one nt from 5′ to 3′ along the sequence. The WFSs detected in the 5′UTR of 24 various ferritin mRNAs are summarized in Table 1. Most of these WFS patterns have high $Zscr_e$ values ($>$ 3.0). All WFSs listed in Table 1 include a common structural core of the IRE that consists a small hairpin structure of 23 nt with a conserved bulge and hairpin loops (Hentze *et al.*, 1988).

We select arbitrarily one sequence, human ferritin light chain, from 24 ferritin mRNAs to perform a random test as mentioned above. Using a window of 60 nt we had 24 570 observations of $Zscr_e$ for the thirty corresponding randomly shuffled sequences. The observed probabilities of WFSs with $Zscr_e \geqslant$ 2.5, 3.0, 3.5 and 3.7 are 0.024, 0.009, 0.004, and 0.002, respectively in the random test. The $Zscr_e$ value of the WFS in the natural sequence of human ferritin light chain is 3.77. The high $Zscr_e$ value signifies the uniqueness of the morphology of the folded IRE structure and it can not be expected to occur by chance.

let-7 RNA and its WFS patterns

Small *let*-7 RNA (~21 nt) can regulate the developmental timing of *Caenorhabditis elegans* (*C.elegans*) and the 21-nt *let*-7 RNA of *C.elegans* can be folded to a well-ordered stem-loop structure with a nearby sequence (Pasquinelli *et al.*, 2000). Using the structural character of the small temporal RNA we search for WFS in the genomic sequences containing *let-7* RNA gene of *C.elegans* (accession no. AF274345), *C.briggsae* (accession no. AF210771) and *D.melanogaster* (accession no. AE003659). *D.melanogaster* sequence contains 260 897 nt, of which the region 54 361–56 700 was used.

We first selected *C.elegans* as a test sequence and used a set of windows from 60 to 95 nt by a step of 5 nt. $Zscr_e$ values were computed by moving these windows in steps of 3 nt from 5′ to 3′ along the sequence. Our results are displayed in Figure 3. In this extensive search we detected 7 WFSs by seven windows from 65 to 95 nt. The WFSs ($Zscr_e \geqslant 6.31$) are segments 1762–1826, 1759–1828, 1756–1830, 1756–1835, 1753–1837, 1750–1839 and 1747–1841. They all occurred in the coincident locations and included the 21-nt *let*-7 RNA (1763–1783). The global maximum (7.74) of $Zscr_e$ values was achieved by the fixed-length window of 75 nt in the segment 1756–1830. Similarly, we also detected the WFS related to the *let*-7 RNA gene in *C.briggsae* and *D.melanogaster*. The distribution of $Zscr_e$ scores computed by a window of 75 nt for the three *let*-7 RNA gene region sequences are shown in Figure 4. The WFSs are segment 1882–1956 in *C.briggsae* and 54 910–54 984 in *D.melanogaster*, and their locations of let-7 RNAs are 1888–1908 and 54 909-54 929, respectively. Both WFSs include the *let*-7 RNA and the corresponding $Zscr_e$ values are 6.22 and 4.99, respectively.

Figure 5 graphically displays the observed distribution of the $Zscr_e$ scores computed in the wild type (*C.elegans*) sequence and 30 randomly shuffled sequences with the same base composition and sequence length. The distinct WFS 1756–1830 ($Zscr_e = 7.74$) can be clearly distinguished from the bulk distribution. The $Zscr_e$ ranged from -1.37 to 5.93 in the 30 random sequences that include 23 880 observations of $Zscr_e$ in the random

Table 1. Functional structural core of iron regulatory elements (IRE) and corresponding WFSs determined in ferritin mRNAs

Ferritin mRNA	Accession No.	mRNA Length (bp)	5′-UTR Size (bp)	Region of IRE Core	Region	WFS Zscr	Sigscr
Human H-chain	L20941	1198	208	31–58	17–71	3.37	−1.90
Human L-chain	BC004245	878	188	17–43	1–60	3.77	−3.22
Rat L-chain	J02741	898	200	32–58	10–74	4.29	−2.11
Mouse H-chain	NM_010239	866	167	30–57	21–65	5.39	−1.98
Mouse L-chain	BC019840	918	187	11–41	1–50	1.51	0.03
Hamster H-chain	M99692	843	139	8–35	1–45	3.09	−1.08
Pig H-chain	D15071	821	112	14–41	6–50	2.85	−1.56
Guinea Pig H-chain	AB073371	1053	126	6–33	2–41	2.92	0.98
Dog	AF285177	816	181	3–32	10–39	2.12	0.18
Chicken H-chain	Y14698	905	151	31–62	12–81	6.29	−1.74
Bullfrog H-chain	M12120	854	141	24–53	9–68	2.89	−1.75
Xenopus	S64727	1388	629	502–531	489–548	4.94	−2.00
Crayfish	X90566	1013	141	10–37	9–38	2.57	0.37
Salmon M-chain	S77386	1010	191	27–58	1–60	3.26	−1.88
Salmon H-chain	AF338763	971	197	11–40	8–47	2.30	−1.10
Rainbow trout H1-chain	D86625	908	168	5–36	1–45	2.31	−1.57
Zebrafish H-chain	AF295373	1062	214	9–32	4-38	2.54	−0.71
Moth L-chain	AF142340	1353	176	63–89	61–95	4.46	−2.21
Butterfly G	AF161710	1330	209	103–129	93–142	5.88	−3.13
Butterfly S	AF161708	1212	211	96–126	93–127	5.12	−2.88
D.melangaster	Y15629	1357	328	151–177	150–179	4.73	−1.79
N.lugens	AJ251148	1853	261	136–162	134–163	5.41	−2.31
Mosquito	L37082	1257	203	85–111	82–116	4.78	−1.92
M.sexta	L47123	1227	134	42–68	34–78	5.29	−2.35

The location of known IRE core in each sequence is listed in column 5. $Zscr_e$ listed in the seventh column was computed from the optimized window whose size that was determined by an extensive search in which the window size was systematically changed from 25 to 80 nt in steps of 5 nt (details see text). Regions of the best $Zscr_e$ in the 5′UTR are listed in column 6. Significance scores (Sigscr) listed in the eighth column were computed by program SIGSTB (Le *et al.*, 1990b). Sigscr is also a standard z-score and computed by $Sigscr = (E - E_r)/std_r$, where E is the lowest free energy computed from a local segment in the sequence. E_r is the sample mean and std_r is the sample standard deviation of the lowest free energies computed from folding 100 randomly shuffled segments of the same size and same base composition as the local segment. In computing Sigscr we scanned the same window with same parameters that were used in computing $Zscr_e$. In general, a local segment with $Sigscr < -3.0$ is considered to be an unusual folding region whose thermodynamic stability is significantly more stable than that by chance (Le and Maizel, 1989; Le *et al.*, 1990b)

sequence of 73 800 nt. Only 6 out of 23 880 observations of $Zscr_e$ are greater than 4.99. The observed probabilities of WFS with $Zscr_e \geqslant 4.0, 3.5, 3.0$ and 2.5 are less than 0.003, 0.006, 0.013 and 0.026 in the random test, respectively. Also, our random tests indicate that $Zscr_e$ is not closely dependent on the GC composition of the segment as shown in Figure 5.

Our results strongly suggest that the WFS detected in *C.elegans* is statistically significant and the well-ordered structure can not be expected by chance. We get a similar conclusion from the random tests for 30 randomly shuffled sequences of *C.briggsae*. Our data show that the $Zscr_e$ scores are ranged from -1.03 to 5.01. There is no WFS with $Zscr_e \geqslant 5.1$ in the random sequences. Only 2 out of 21 930 observations of $Zscr_e$ are equal to or greater than 4.99. The test for 30 random sequences of *D.melanogaster* shows that 10 out of 22 680 observations of $Zscr_e$ in the random sequences of 70 200 nt are equal to or greater than 4.99. The observed probability of such distinct WFS with $Zscr_e \geqslant 4.99$ is small and about 0.0004 in the random test.

Draper (1996) has suggested that RNA functional elements whose folded structure conformation plays a crucial role are known to possess a specific structural feature that is both thermodynamically stable and uniquely folded. In this study the conformational property in the RNA secondary structure is measured by the quantitative score, $Zscr_e$, that is determined based on the consideration of both thermodynamic stability and uniqueness of the minimal free energy structure. Our computational results show that WFSs with high $Zscr_e$ are coincident with the reported functional structure RNA elements in our tested examples and are unlikely to occur by chance. Statistical extremes with high $Zscr_e$ values determined by our method represent such significant folding segments where predicted structures are expected to be well-ordered.

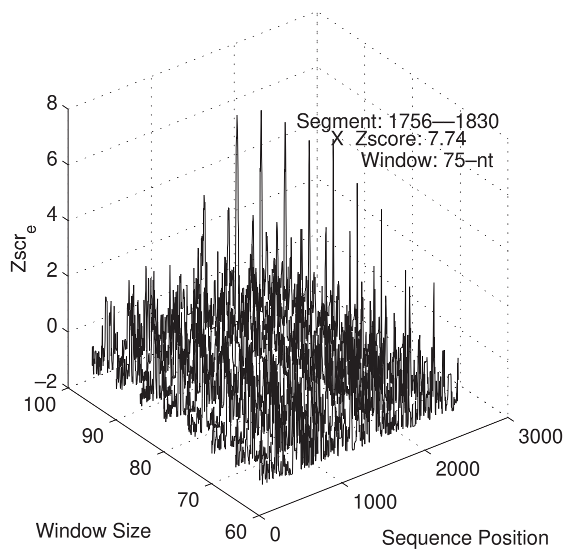

Fig. 3. A 3D plot of $Zscr_e$, the start position of a local segment and the window size. The $Zscr_e$ values were computed by moving a fixed-length window in steps of 3 nt from $5'$ to $3'$ along the sequence of *C.elegans let-7* RNA gene region. In the computation, the window size was systematically changed from 60 to 95 nt by a step of 5 nt. The maximum $Zscr_e$ (= 7.74) corresponds to the WFS 1756–1830 that was detected by a window of 75 nt and marked in the plot.

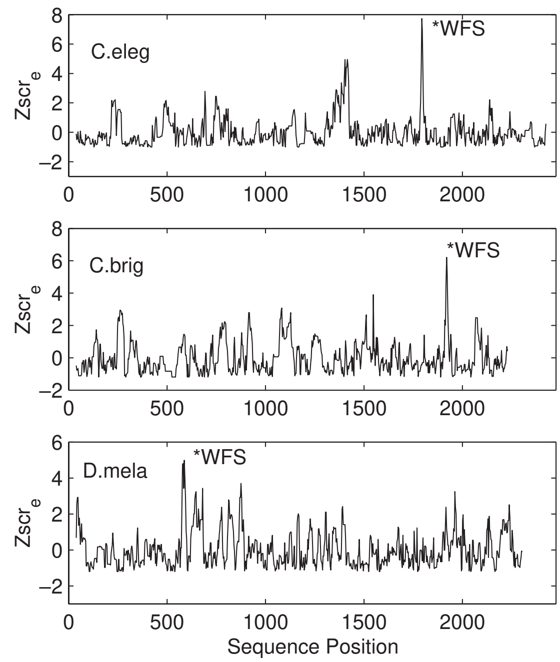

Fig. 4. $Zscr_e$ distributions of the three genomic sequences of *C.elegans* (top), *C.briggsae* (middle) and *D.melanogaster* (bottom). Each plot was produced by plotting $Zscr_e$ against the position of the middle base in the window of 75 nt. The detected WFSs including the 21-nt let-7 RNA are asterisked. In the plot of *D.melanogaster*, the sequence position 54 361 is numbered as position 1.

It has been reported that 'well-determined' structural domains of ribosomal RNAs can be predicted and observed in 'energy dot plots' (Zuker and Jacobson, 1995). Recently, Schultes *et al.* (1999) proposed quantitative measures to estimate the stability and uniqueness of RNA secondary structures based on the mean length of stems and total number of base pairs in the predicted RNA secondary structures from RNAfold (Zuker and Steigler, 1981) and/or VIENNA (Hofacker *et al.*, 1994). The distinct conformation in the structure was not considered throughly in their methods. We previously reported a computational method to discover structured elements in mRNA sequences (Le and Maizel, 1989; Le *et al.*, 1990b). The program SEGFOLD and its modified version SIGSTB are used to explore a sequence by choosing successive segments of an mRNA and comparing the energy difference computed between the natural sequence and a number of randomly shuffled sequences of the same length and base composition. Simultaneously, we also compare the lowest free energy of each segment with the average energy of all segments of same size in the RNA sequence. The quantitative measures, significance and stability scores (Sigscr and Stbscr) used in SEGFOLD and SIGSTB are based on the thermodynamic stability only. The new method puts our emphasis on the uniqueness of the morphology of the folded RNA secondary structure. In general both methods provide similar results for the discovery of the 'well-determined' structures with high stability as we observed in the case of HIV-1 mRNA. But our computational experiment indicates that the well-ordered structure of ferritin IRE is only moderately stable. The methods of SEGFOLD and SIGSTB fail to discover the distinct IRE for most of 24 ferritin mRNAs. However, the functional IRE elements can be detected by our new method ed_scan. The results presented in this paper demonstrate that the uniqueness of the folded RNA structures is represented by the new measure $Zscr_e$. However, it should be noted that the comparison of thermodynamic stability used in ed_scan are limited. In the comparison, we consider only two extreme cases of the lowest free energy structure and its corresponding ORS. This is not sufficient to characterize all distinct conformations for the natural structure. Also, this method does not detect the class of RNA sequences where the minimal energy of predicted structure in the natural sequence is less stable than its corresponding random structures. SEGFOLD and SIGSTB are able

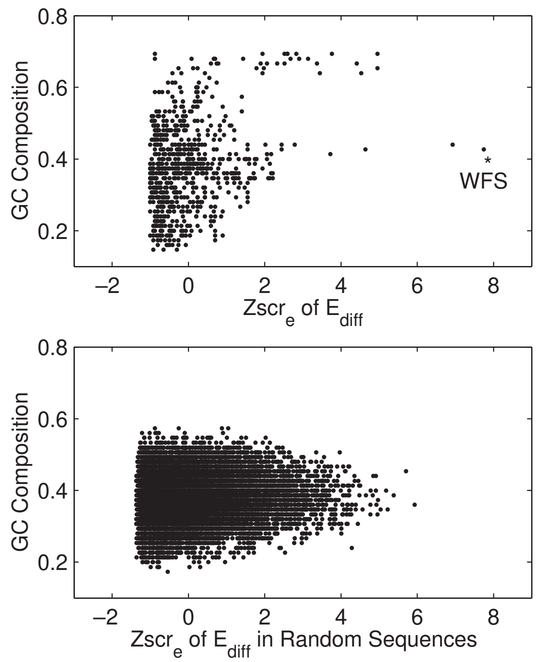

Fig. 5. Relationships between GC composition and $Zscr_e$ of local segments of 75 nt computed in wild type sequence of *C.elegans* (top) and its 30 randomly shuffled sequences (bottom). $Zscr_e$ values were computed by moving a 75-nt window in steps of 3 nt from 5′ to 3′ along the sequences. The WFS with high $Zscr_e$ is apparently separated from the bulk distribution in the positive direction.

to find such features (Le *et al.*, 1989) and thus provide valuable complementary tools to ed_scan. Nevertheless our method is helpful in the determination of local RNA elements with structure-dependent functions in mRNAs. This is especially applicable to knowledge discovery in the post-genomic age.

CONCLUSION

We have proposed a computational method to search for the local WFS with high $Zscr_e$ in a nucleotide sequence. The WFS is measured by the difference of free energies between the optimized and its corresponding ORS. Our computational experiments for native and random sequences of HIV-1 mRNA, 24 ferritin mRNAs and three *let*-7 RNA gene region sequences show that the proposed $Zscr_e$ is a good measure to estimate quantitatively the significance of the well-ordered conformation for local RNA structures in the sequence. Our results indicate that significant WFSs are distinguishable from the bulk distribution and are closely associated with the known functional RRE, IRE and small temporal *let*-7 RNA. Our method is helpful in the determination of potential functional elements with structure dependent functions in nucleic acid sequences.

ACKNOWLEDGEMENTS

This content of this publication do 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 US Government.

REFERENCES

Bashirullah,A., Cooperstock,R.L. and Lipshitz,H.D. (1998) RNA localization in development. *Annu. Rev. Biochem.*, **67**, 335–394.

Chen,S.-J. and Dill,K.A. (2000) RNA folding energy landscapes. *Proc. Natl Acad. Sci. USA*, **97**, 646–651.

Dayton,E.T., Konings,D.A., Powell,D.M., Shapiro,B.A., Butini,L., Maizel,Jr,J.V. and Dayton,A.I. (1992) Extensive sequence-specific information throughout the CAR/RRE, the target sequence of the human immunodeficiency virus type 1 Rev protein. *J. Virol.*, **66**, 1139–1151.

Draper,D.E. (1996) Strategies for RNA folding. *Trends Biochem. Sci.*, **21**, 145–149.

Forsdyke,D.R. (1995) Conservation of stem-loop potential in introns of snake venom phospholipase A genes. An application of FORS-D analysis. *Mol. Biol. Evol.*, **12**, 1157–1165.

Gray,N.K. and Wickens,M. (1998) Control of translation initiation in animals. *Annu. Rev. Cell Dev. Biol.*, **14**, 399–458.

Hentze,M.W., Caughman,S.W., Casey,J.L., Koeller,D.M., Rouault,T.A., Harford,J.B. and Klausner,R.D. (1988) A model for the structure and functions of iron-responsive elements. *Gene*, **72**, 201–208.

Hermann,T. and Patel,D.J. (1999) Stitching together RNA tertiary architectures. *J. Mol. Biol.*, **294**, 829–849.

Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte Chem.*, **125**, 167–188.

Le,S.-Y. and Maizel,Jr,J.V. (1989) A method for assessing the statistical significance of RNA folding. *J. Theor. Biol.*, **138**, 495–510.

Le,S.-Y., Chen,JihH., Chatterjee,D. and Maizel,Jr,J.V. (1989) Sequence divergence and open regions of RNA secondary structures in the envelope regions of the 17 human immunodeficiency virus isolates. *Nucleic Acids Res.*, **17**, 3275–3288.

Le,S.-Y., Malim,M.H., Cullen,B.R. and Maizel,Jr,J.V. (1990a) A highly conserved RNA folding region coincident with the Rev response element of primate immunodeficiency viruses. *Nucleic Acids Res.*, **18**, 1613–1623.

Le,S.Y., Chen,J.H. and Maizel,Jr,J.V. (1990b) Efficient searches for unusual folding regions in RNA sequences. In Sarma,R.H. and Sarma,M.H. (eds), *Structure and Methods: Human Genome Initiative and DNA Recombination, I.* Adenine Press, Schenectady, NY, pp. 127–136.

Malim,M.H., Hauber,J., Le,S.-Y., Maizel,Jr,J.V. and Cullen,B.R. (1989) The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature*, **338**, 254–257.

Mann,D.A., Mikaelian,I., Zemmel,R.W., Green,S.M., Lowe,A.D., Kimura,T., Singh,M., Butler,P.J., Gait,M.J. and Karn,J. (1994) A molecular rheostat: co-operative rev binding to stem of the rev-response element modulates human immunodeficiency virus type-1 late gene expression. *J. Mol. Biol.*, **241**, 193–207.

Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.

Pasquinelli,A.E., Reinhart,B.J., Slack,F., Martindale,M.Q., Kuroda,M.I., Maller,B., Hayward,D.C., Ball,E.E., Degnan,B., Muller,P. *et al.* (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, **408**, 86–88.

Patzel,V. and Sczakiel,G. (1997) The hepatitis B virus posttranscriptional regulatory element contains a highly stable RNA secondary structure. *Biochem. Biophys. Res. Commun.*, **231**, 864–867.

Rivas,E. and Eddy,S.R. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–605.

SantaLucia,Jr,J. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.

Schultes,E.A., Hraber,P.T. and LaBean,T.H. (1999) Estimating the contributions of selection and self-organization in RNA secondary structure. *J. Mol. Evol.*, **49**, 76–83.

Simons,R.W. and Grunberg-Manago,M. (eds) (1998) *RNA Structure and Function*. Cold Spring Harbor Press, New York.

Tinoco,Jr,I. and Bustamante,C. (1999) How RNA folds. *J. Mol. Biol.*, **293**, 271–281.

Zuker,M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.

Zuker,M. and Steigler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–149.

Zuker,M. and Jacobson,A.B. (1995) ‘Well-determined’ regions in RNA secondary structure prediction: analysis of small subunit ribosomal RNA. *Nucleic Acids Res.*, **23**, 2791–2798.