3574–3582 Nucleic Acids Research, 2002, Vol. 30 No. 16

RNA molecules with structure dependent functions are uniquely folded

Shu-Yun Le*, Kaizhong Zhang[1] and Jacob V. Maizel Jr

Laboratory of Experimental and Computational Biology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, Building 469, Room 151, Frederick, MD 21702, USA and [1]Department of Computer Science, University of Western Ontario, London, ON N6A 5B7, Canada

Received April 22, 2002; Revised and Accepted June 18, 2002

ABSTRACT

***Cis*-acting elements in post-transcriptional regulation of gene expression are often correlated with distinct local RNA secondary structure. These structures are expected to be significantly more ordered than those anticipated at random because of evolutionary constraints and intrinsic structural properties. In this study, we introduce a computing method to calculate two quantitative measures, *NRd* and *Stscr*, for estimating the uniqueness of an RNA secondary structure. *NRd* is a normalized score based on evaluating how different a natural RNA structure is from those predicted for its randomly shuffled variants. The lower the score *NRd* the more well ordered is the natural RNA structure. The statistical significance of *NRd* compared with that computed from structural comparisons among large numbers of randomly permuted sequences is represented by a standardized score, *Stscr*. We tested the method on the *trans*-activation response element and Rev response element of HIV-1 mRNA, internal ribosome entry sequence of hepatitis C virus, *Tetrahymena thermophila* rRNA intron, 100 tRNAs and 14 RNase P RNAs. Our data indicate that functional RNA structures have high *Stscr*, while other structures have low *Stscr*. We conclude that RNA functional molecules and/or *cis*-acting elements with structure dependent functions possess well ordered conformations and they are uniquely folded as measured by this technique.**

INTRODUCTION

Numerous experimental results have shown that RNA molecules perform a wide range of functions in biological systems. The known biological functions of RNAs continue to grow and its important role in the regulation control of gene expression is evident in many different biological fields (1,2). Though single-stranded regions exist in most RNAs, distinct well ordered structure in local segments of single-stranded RNA sequences often correlates with functions such as control of replication, transcription, mRNA processing, translation and metabolism (1–4) making it desirable to understand the conformation of associated local RNA structures.

Some combinations of base pairings in stem–loops and some distinct loop sequences are more abundant in functional structural RNAs (FSRs) (5–8). Therefore, the prediction of distinct folding patterns in RNA sequences is an important goal of genomic sequence analysis. It has been suggested that the FSRs possess well ordered conformations that are both thermodynamically stable and uniquely folded (9,10). This is because the functional elements must be optimized both in the conformational properties and sequence patterns where the interactions between RNA and RNA, as well as RNA and protein, play a crucial role in their functions (11). In the study of energy landscapes of RNA folding, we have demonstrated that the FSR elements are often thermodynamically more stable than anticipated in their corresponding random structures (12,13). This implies that evolution has constrained the structural properties of FSR elements to be thermodynamically stable. However, no available methods have been developed for estimating the structural uniqueness of the well ordered RNA structures based on the structural comparison in detail.

There are a few RNA motifs whose three-dimensional structures are known. It is also true that even random RNA sequences can be folded so that their complementary sequences form double helical stems. Functional RNA molecules found in modern organisms are evolutionary products (11). Evolution would increase the thermodynamic stability of the folded structure and reduce the possibility of alternative folding forms. Although we do not fully understand how to measure the structural uniqueness of distinct RNA structures accurately, it is reasonable to suppose FSRs have evolved morphologies or distinct conformations that are not expected to be found by chance. We expect that the more different a natural evolved RNA structure is from a large number of random and unevolved structures, the more unique is the natural RNA structure.

In this study we present a novel method to determine the uniqueness of well ordered RNA secondary structures. We test 100 tRNA molecules, 14 ribonuclease P (RNase P) RNA molecules, *Tetrahymena thermophila* rRNA intron, internal ribosome entry sequence (IRES) of hepatitis C virus, *trans*-activation response (TAR) element and Rev response element (RRE) of HIV-1. We find that most RNA structures derived by

*To whom correspondence should be addressed. Tel: +1 301 846 5576; Fax: +1 301 846 5598; Email: shuyun@orleans.ncifcrf.gov

phylogenetic methods and some structures from minimal energy dynamic programming algorithm have well ordered conformations.

MATERIALS AND METHODS

Quantitative measures for the uniqueness of RNA structures

It has been suggested that RNA folding is hierarchical and sequential (14). The primary sequence of a natural RNA determines its secondary structure and the secondary structure determines its tertiary structure. Since RNA secondary structure is supposed to change slightly by the additional tertiary interactions (14), it is a crucial step to characterize the RNA secondary structure for our understanding of the uniqueness of the RNA three-dimensional structure. In this study we focus our attention on the uniqueness of secondary structures folded by RNA molecules.

To define the uniqueness of RNA secondary structures quantitatively we hypothesize that the well ordered conformation of functional RNA molecules is expected to be rare in the conformation space formed from a population of random sequences with the same base compositions and same length as the sequence of natural RNAs. For a given natural RNA sequence, we generate 300 randomly shuffled sequences. In the random shuffling, nucleotides at all sites of the natural sequence are sequentially swapped with a randomly chosen site elsewhere in the sequence (15). If the secondary structure of the natural RNA was established experimentally or by phylogenetic comparisons, the established structure is used. Otherwise, the lowest free energy structure predicted by Zuker's mfold with Turner energy rules (16,17) is used. The secondary structures of the random sequences (termed random structures) are also assumed to be folded with the lowest free energy computed by mfold. In consideration of the fact that the predicted optimized structure by mfold does not include any tertiary interactions, the tertiary interactions conserved in the inferred structure by phylogenetic comparisons are not taken into account in the structure comparison.

In order to facilitate the structural comparison, we define a maximal matching score (MMS) to represent the maximal structural similarity between two RNA structures. The MMS is computed by the program rna_match (see below). Therefore we have 300 MMS observations computed between the structure of the natural RNA and the 300 random structures. The sample mean of the 300 MMS observations is termed as *NR*. A quantitative measure of the uniqueness of RNA secondary structures is then defined as the average *NR* per nucleotide, which is a normalized score and denoted by *NRd*. It represents the density of MMS between the structure of a natural RNA and a large number of the random structures. *NRd* is taken as a measure of the uniqueness of the structural morphology folded by an RNA molecule. The lower the *NRd*, the more unique is the well ordered structure of an RNA molecule.

What is the average of MMS values between any two random structures? Is the score *NRd* (or *NR*) computed from a natural RNA molecule statistically significant? To address these questions, we perform a statistical test for a set of randomly shuffled sequences by the same procedure as we used in the computation of *NR*. We generally collect a set of 25 random sequences that are also generated by random permutations as mentioned above. For each of the 25 random sequences we compute the MMS repeatedly by comparing the folded structure with the each of the 300 random structures as used previously. As a result, we have 25×300 observations of MMS for these random structures. The sample mean and sample standard deviation of the 7500 random versus random MMS observations are termed as *RR* and *std*, respectively. Consequently, we can define *Stscr* as a significance score of the uniqueness of RNA secondary structures. The *Stscr* is defined as

$$Stscr = (RR - NR)/std.$$

The greater *Stscr*, the statistically more significant is the well ordered structure of the natural RNA. If score *Stscr* is treated as a normal variable with zero mean and unit standard deviation, the significance level achieved can be determined by the normal distribution. Thus, the probability of observing a departure from the mean of >1.96 standard deviations ($Stscr \geqslant 1.96$) is 0.05.

Computing similarity between two RNA structures

In our approach (18,19) of RNA molecule comparison, we define three edit operations, substitute, delete and insert. For a given RNA molecule, each operation can be applied to either a base pair or an unpaired base. When applied to unpaired bases, these operations are exactly the same as in sequence matching. To substitute paired bases is to replace one base pair with another. This means that at the sequence level, two bases are changed at the same time. To delete a base pair is to delete the two bases of the base pair. At the sequence level, this means to delete two bases at the same time. To insert a base pair is to insert a new base pair. At the sequence level, this means to insert two bases at the same time. Note that there is no substitute operation that can change a base pair to an unpaired base or vice versa.

We assume that there is a score function associated with the operations. The score function for base pairs is defined on $\Sigma \times \Sigma \cup \{\lambda\}$, and the score function for unpaired bases is defined on $\Sigma \cup \{\lambda\}$ (20). With these definitions, we can consider how to translate one RNA into another using the optimal number of weighted operations. With appropriate score functions, this can give us either a similarity or a distance measure between two RNA structures. We take into account both the sequence and the structural information of RNA molecules. Our measure treats a base pair as a unit and does not allow it to match to two unpaired bases. This is closer to the spirit of the comparative analysis method currently used in the analysis of RNA secondary structures. When one base of a base pair changes, we usually find that its complementary base also changes so as to conserve that base pair in RNA structures.

The algorithm (18–20) of structure comparison (rna_match) is briefly described here. For any RNA R, we use $r[i]$ to represent the ith base in R and we use $R[i \cdots j]$ to represent the substructure formed by bases from $r[i]$ to $r[j]$. Let $R_1[1 \cdots m]$ and $R_2[1 \cdots n]$ be the two given RNA structures. We use $M(i_1, i_2; j_1, j_2)$ to represent MMS between $R[i_1 \cdots i_2]$ and $R_2[j_1 \cdots j_2]$. Suppose that we want to compute the MMS 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 it is clear that

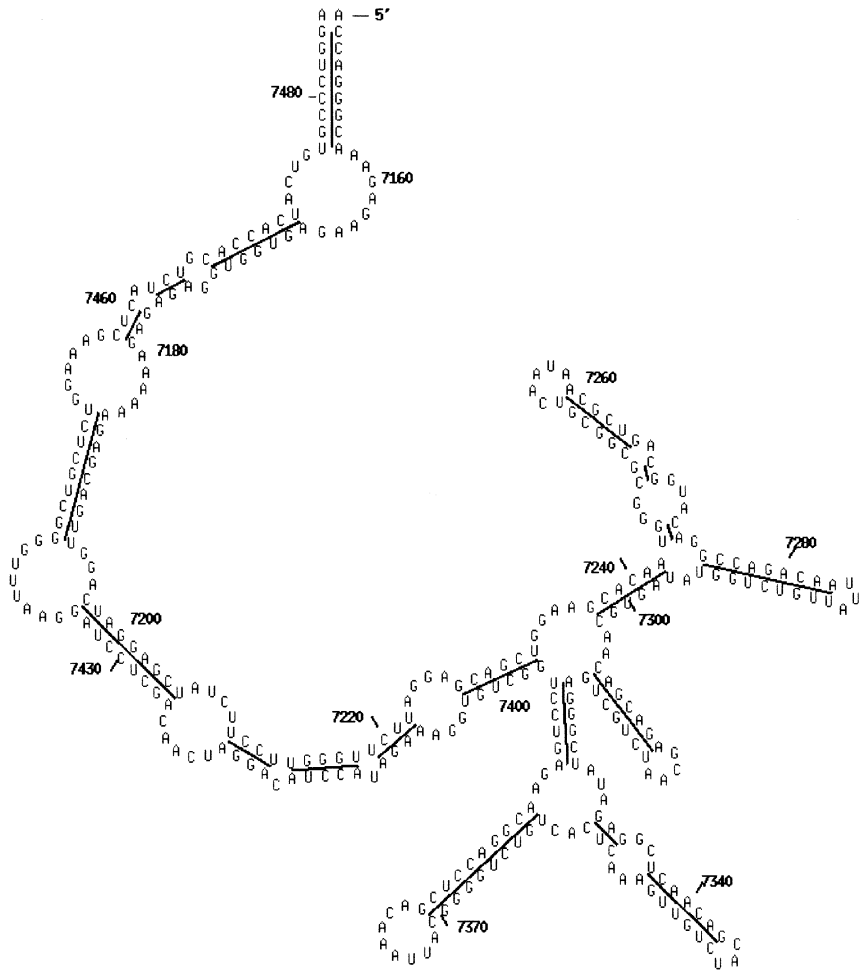

Figure 1. Conserved secondary structure of RRE sequence of HIV-1. The RRE sequence is located at the region of 7149–7485 of the mRNA of HIV-1 (isolate U455).

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

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 substitution, 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 can show 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.

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

The above are the two main cases and the other cases can be handled similarly. Because of the second formula, we know that in order to compute MMS between R_1 and R_2, $M(1, m; 1, n)$, we need to have MMS between some substructures, namely $M(i'+1, i-1; j'+1, j-1)$ where $(r_1[i'], r_1[i])$ and $(r_2[j'], r_2[j])$ are both base pairs, available. This suggests a bottom-up dynamic programming algorithm to find MMS between R_1 and R_2. We consider the smaller substructures first and eventually consider the whole structures R_1 and R_2. The worst case time complexity of the algorithm is $O(m \times n \times stem(R_1) \times stem(R_2))$, where $stem(R_1)$ and $stem(R_2)$ are the number of stems in R_1 and R_2, respectively. In practice the time complexity is roughly $O(m \times n \times \log(m) \times \log(n))$. The space complexity of the algorithm is $O(m \times n)$.

RESULTS

In our computational experiments, we tested the sequences of 100 tRNAs that had been used by Eddy and Durbin (21). The secondary structures of cloverleaf models of the 100 tRNAs were extracted from the tRNA database (22). The common secondary structure models derived by phylogenetically comparative analysis are all well established for RNase P RNAs (23) and group I introns (24). The conserved secondary

Table 1. Significance scores (*Stscr*) of the uniqueness of RNA secondary structures from 100 tRNA molecules

Sequences		RNA Secondary Structures Computed from Natural tRNA Sequences						Randomly Permuted Sequences		
tRNAs	Size	Optimized Structures			Phylogenetic Structures					
Code	(nt)	Stscr	NR	NRd	Stscr	NR	NRd	RR	RRd	std
DA2480	76	2.93	45.54	0.60	2.93	45.54	0.60	78.34	1.03	11.20
DA5280	69	1.16	48.63	0.70	2.12	29.13	0.42	72.12	1.05	20.28
DA7680	73	3.16	34.39	0.47	3.16	34.39	0.47	76.51	1.05	13.28
DC0380	76	2.40	36.90	0.49	2.40	36.90	0.49	79.37	1.04	17.73
DC5020	69	-0.79	74.77	1.08	3.60	21.45	0.31	65.14	0.94	12.15
DD2680	73	-0.53	82.66	1.13	2.48	29.36	0.40	73.27	1.00	17.72
DD2920	74	-0.73	89.99	1.22	2.11	28.93	0.39	74.35	1.00	21.55
DD4000	72	1.99	49.97	0.69	3.49	29.62	0.41	77.14	1.07	13.63
DD4080	72	0.55	62.33	0.87	2.19	28.05	0.39	73.73	1.02	20.90
DD5280	68	2.05	41.99	0.62	2.05	41.99	0.62	66.98	0.99	12.19
DD5320	70	2.02	41.21	0.59	2.71	32.61	0.47	66.43	0.95	12.46
DE1230	76	0.56	57.67	0.76	1.82	38.99	0.51	65.99	0.87	14.84
DE1660	76	-0.68	84.87	1.12	2.88	31.90	0.42	74.78	0.98	14.87
DE6160	72	2.62	19.01	0.26	2.62	19.01	0.26	64.43	0.89	17.31
DF1180	76	-1.06	93.18	1.23	2.65	32.62	0.43	75.85	1.00	16.31
DF5220	70	2.83	36.39	0.52	2.73	37.10	0.53	56.61	0.81	7.15
DF5900	68	-0.27	56.81	0.84	2.14	29.19	0.43	53.76	0.79	11.50
DF5930	68	2.46	31.91	0.47	2.46	31.91	0.47	49.60	0.73	7.18
DF9160	74	1.49	44.80	0.61	2.55	28.39	0.38	67.91	0.92	15.48
DG2440	73	0.58	57.35	0.79	2.19	27.84	0.38	68.01	0.93	18.36
DG2921	71	2.31	47.67	0.67	2.61	44.61	0.63	70.84	1.00	10.04
DG4070	72	0.29	54.35	0.75	2.09	34.16	0.47	57.64	0.80	11.25
DG5040	69	1.61	41.94	0.61	2.20	35.47	0.51	59.67	0.86	11.01
DG7740	71	-0.53	94.42	1.33	3.19	27.12	0.38	84.85	1.20	18.09
DG8100	72	0.37	66.84	0.93	1.98	45.08	0.63	71.88	1.00	13.53
DH1700	77	1.97	36.97	0.48	1.97	36.97	0.48	75.11	0.98	19.39
DH2880	76	-0.10	74.57	0.98	2.01	46.81	0.62	73.19	0.96	13.11
DH2960	76	-0.13	76.77	1.01	3.00	41.13	0.54	75.34	0.99	11.41
DH4360	75	1.07	55.19	0.74	2.24	34.52	0.46	74.12	0.99	17.68
DH5120	68	0.09	49.09	0.72	2.55	-12.75	-0.19	51.43	0.76	25.17
DI2220	74	2.79	20.09	0.27	2.79	20.09	0.27	70.09	0.95	17.89
DI2701	72	2.12	40.13	0.56	2.12	40.13	0.56	75.39	1.05	16.63
DI2400	74	4.42	24.17	0.33	4.42	24.17	0.33	78.99	1.07	12.41
DK1230	76	0.01	70.44	0.93	2.56	28.69	0.38	70.57	0.93	16.29
DK4340	73	-0.87	86.20	1.18	2.69	38.57	0.53	74.56	1.02	13.37
DK5320	65	-0.86	63.11	0.97	0.78	45.56	0.70	53.91	0.83	10.73
DK6280	73	2.02	7.78	0.11	2.02	7.78	0.11	64.70	0.89	28.22
DL0980	84	-0.85	102.11	1.22	2.62	27.52	0.33	83.88	1.00	21.47
DL1141	83	-1.17	90.75	1.09	5.06	1.99	0.02	74.07	0.89	14.26
DL1200	86	3.36	9.31	0.11	3.36	9.31	0.11	75.68	0.88	19.76
DL1231	85	-1.15	94.87	1.12	3.09	13.43	0.16	72.83	0.86	19.24
DL1543	82	0.49	72.80	0.89	4.49	10.71	0.13	80.48	0.98	15.55
DL1662	85	0.99	54.73	0.64	3.49	17.14	0.20	67.93	0.80	13.40
DL1750	86	-0.21	84.23	0.98	4.99	7.79	0.09	81.14	0.94	14.70
DL2522	79	0.45	67.17	0.85	4.36	-11.87	-0.15	76.20	0.96	20.18
DL3200	84	-0.81	87.07	1.04	2.75	27.07	0.32	73.43	0.87	16.83
DL4070	83	-0.12	75.41	0.91	3.88	9.78	0.12	73.39	0.88	16.37
DL4700	66	1.48	42.86	0.65	2.03	35.87	0.54	61.41	0.93	12.56
DL4760	67	2.15	46.11	0.69	2.15	46.11	0.69	77.91	1.16	14.79

structures of TAR (25) of HIV-1 and IRES of HCV (26) are from previous publications. The phylogenetically conserved secondary structure of RRE of HIV-1 (isolate U455) is derived from 151 HIV-1 sequences (Fig. 1).

tRNA molecules possess well ordered conformations

The results detailed in Table 1 (columns 6–8) demonstrate that the MMS between the classical cloverleaf structure of natural tRNA derived from phylogenetic and experimental means and corresponding random structures is significantly less than MMS for random versus random sequences in general. Only 10 out of 100 tRNAs have the density of MMS, $NRd \geqslant 0.60$. The calculated *NRd* ranges from −0.19 to 0.70 in the 100 tested tRNA molecules with an average of 0.38 ± 0.18. The calculated *RRd* among permuted random sequences of the tRNAs are ranged from 0.73 to 1.22 and averaged to 0.96 ± 0.11. The average *RRd* from random sequences is ~5.3 standard deviations higher than the average *NRd*

Table 1. *Continued*

Sequences		RNA Secondary Structures Computed from Natural tRNA Sequences						Randomly Permuted Sequences		
tRNAs	Size	Optimized Structures			Phylogenetic Structures					
Code	(nt)	Stscr	NR	NRd	Stscr	NR	NRd	RR	RRd	std
DL5880	71	-0.36	60.18	0.85	2.76	33.89	0.48	57.13	0.80	8.43
DL7740	83	-1.23	108.47	1.31	3.07	19.67	0.24	83.01	1.00	20.62
DL7920	83	1.33	49.31	0.59	4.28	1.31	0.02	70.97	0.86	16.26
DM4000	74	-0.96	81.83	1.11	2.39	33.29	0.45	67.92	0.92	14.52
DN4620	70	2.46	61.31	0.88	7.59	19.45	0.28	81.42	1.16	8.16
DP0680	75	2.90	31.38	0.42	3.00	29.53	0.39	83.15	1.11	17.85
DP1360	74	3.03	40.33	0.54	3.03	40.33	0.54	85.85	1.16	15.03
DQ1340	73	-0.78	88.39	1.21	2.05	34.14	0.47	73.45	1.01	19.16
DQ3220	72	-0.03	87.34	1.21	3.97	37.81	0.53	86.92	1.21	12.36
DQ4880	69	0.15	78.86	1.14	3.05	16.20	0.23	82.14	1.19	21.63
DQ5080	71	0.78	50.00	0.70	2.03	31.56	0.44	61.66	0.87	14.86
DQ6160	73	2.77	32.13	0.44	2.77	32.13	0.44	74.34	1.02	15.25
DR1660	77	-0.46	93.23	1.21	3.85	30.14	0.39	86.52	1.12	14.65
DR1663	77	-0.95	89.47	1.16	2.47	34.79	0.45	74.29	0.96	15.97
DR3320	74	0.21	64.89	0.88	2.38	20.55	0.28	69.20	0.94	20.44
DR5080	69	0.38	55.69	0.81	2.03	33.09	0.48	60.83	0.88	13.44
DR6051	73	-0.26	81.69	1.12	3.20	31.31	0.43	77.91	1.07	14.54
DS0261	94	0.34	71.02	0.76	5.22	15.59	0.17	74.89	0.80	11.37
DS1230	91	2.75	16.52	0.18	3.46	1.47	0.02	74.72	0.82	21.14
DS2480	87	2.84	30.34	0.35	3.12	25.43	0.29	80.30	0.92	17.56
DS2520	87	1.22	46.93	0.54	3.07	17.12	0.20	66.59	0.77	16.11
DS2922	92	-1.00	93.46	1.02	3.76	27.80	0.30	79.68	0.87	13.78
DS6745	82	-0.22	91.43	1.11	6.67	-0.70	-0.01	88.51	1.08	13.38
DT0661	74	2.50	46.88	0.63	2.50	46.88	0.63	81.20	1.10	13.75
DT1542	75	2.60	20.21	0.27	2.60	20.21	0.27	78.35	1.04	22.33
DT2600	72	2.43	30.45	0.42	2.43	30.45	0.42	63.63	0.88	13.66
DT3880	71	1.99	47.43	0.67	1.99	47.43	0.67	68.07	0.96	10.36
DT4700	63	-0.22	64.68	1.03	2.23	36.23	0.58	62.15	0.99	11.63
DT4880	65	0.16	77.81	1.20	4.28	33.62	0.52	79.57	1.22	10.74
DT6160	72	1.68	45.08	0.63	1.54	46.78	0.65	66.47	0.92	12.75
DT7740	74	1.31	51.53	0.70	2.90	24.27	0.33	73.93	1.00	17.14
DT9991	93	2.68	22.98	0.25	2.68	22.98	0.25	71.46	0.77	18.06
DV2600	72	3.27	31.72	0.44	3.27	31.72	0.44	66.29	0.92	10.58
DV3200	72	4.54	31.02	0.43	4.54	31.02	0.43	78.20	1.09	10.40
DV3960	75	-0.23	63.28	0.84	2.83	18.51	0.25	59.86	0.80	14.61
DV4000	71	2.36	35.92	0.51	2.36	35.92	0.51	68.66	0.97	13.90
DW4360	74	3.43	28.49	0.38	2.96	33.93	0.46	68.27	0.92	11.60
DW5080	70	0.11	56.55	0.81	2.72	31.09	0.44	57.64	0.82	9.77
DX0260	80	-1.05	91.21	1.14	4.02	19.88	0.25	76.49	0.96	14.08
DX4320	74	2.96	24.73	0.33	2.96	24.73	0.33	71.59	0.97	15.84
DX4440	74	-0.54	75.21	1.02	2.83	32.86	0.44	68.43	0.92	12.55
DX4960	69	2.57	31.43	0.46	2.46	32.57	0.47	60.06	0.87	11.16
DX5160	68	0.87	51.31	0.75	2.28	39.93	0.59	58.27	0.86	8.03
DY0660	74	2.74	32.90	0.44	2.74	32.90	0.44	77.85	1.05	16.41
DY2920	84	-0.94	101.54	1.21	2.33	5.11	0.06	73.90	0.88	29.54
DY3280	84	-1.33	101.07	1.20	5.15	8.35	0.10	82.03	0.98	14.32
DY3770	81	-0.01	77.07	0.95	3.96	28.36	0.35	76.94	0.95	12.26
DY4880	66	-0.38	74.93	1.14	1.98	35.84	0.54	68.61	1.04	16.53
DY5040	71	2.52	48.54	0.68	2.52	8.54	0.12	66.73	0.94	7.22
DY5220	71	-0.32	81.15	1.14	2.38	34.55	0.49	75.65	1.07	17.25
DY6743	73	2.66	23.39	0.32	2.76	21.54	0.30	76.21	1.04	19.84

indicating that the structural conformations of the natural tRNAs are significantly different from those of corresponding random structures. The significance scores of the uniqueness of the common tRNA secondary structure computed from 100 wild-type tRNAs are high and the *Stscr* values average to 2.94 $\pm$ 1.02. Only 3 out of 100 wild-type tRNAs have *Stscr* values <1.96. It indicates that the uniqueness of the common tRNA secondary structure of cloverleaf representations is statistically significant.

Table 1 (columns 3–5) also shows the results of repeating the procedure described above using the optimized structures, rather than the common cloverleaf structures. Our data show that 26 tRNA sequences are correctly predicted to be the classical cloverleaf structure by mfold with the Turner energy

rules (16,17) and that they all have high *Stscr* values. However, the significance scores *Stscr* of the optimized structures computed from most of the other 74 wild-type tRNAs are much lower and their average value is 0.36 ± 1.22. The average value of *Stscr* of the corresponding 74 cloverleaf structures is 3.02 ± 1.12. It is evident that for tRNAs the structural morphologies of phylogenetically inferred structures are more ordered than those observed in the predicted minimal free energy structures.

Among the 74 tRNAs, three predicted optimized structures have *Stscr* a little higher than that computed by the common cloverleaf structures. The three tRNAs are coded by DW4360, DF5220 and DX4960. The optimized structures of these three tRNAs are very close to the corresponding cloverleaf structures. Four out of the 74 lowest free energy structures have *Stscr* scores >1.96 in which the optimized structures are quite different from the functional cloverleaf structures. They are tRNAs DS1230, DY6743, DN4620 and DG2921. The *Stscr* values computed by the optimized structure are 2.75, 2.66, 2.46 and 2.31, respectively. But the corresponding scores of *Stscr* computed by the common functional structures are 3.46, 2.76, 7.59 and 2.61, which are greater than those computed from the optimized structures. The structural difference between the well ordered conformation of the phylogenetic cloverleaf structures and the predicted minimal energy structures is statistically significant. The bulk distribution of *NRd* of natural tRNA structures is different from the random structures as shown in Figure 2. Also the *Stscr* distribution computed from the classical cloverleaf structures is different and separate from that of the corresponding lowest free energy structures.

RNase P RNAs are also uniquely folded

NRd and *Stscr* of secondary structures of 14 RNase P RNAs are listed in Table 2. *NRd* values from the phylogenetically inferred secondary structures range from 0.04 to 0.31 and average to 0.16 ± 0.08. The *NRd* from the optimized structures predicted by mfold ranges from 0.06 to 0.54 and average to 0.37 ± 0.13. *RRd* calculated from permuted random sequences ranged from 0.38 to 0.61 and averaged to 0.48 ± 0.05. Well ordered conformations in the phylogenetically conserved structures of RNase P RNAs are easily distinguished from those of permuted random sequences by the *NRd* and *RRd*. *Stscr* of 14 phylogenetically inferred structures range from 2.15 to 5.84 and average to 3.73 ± 1.25. All 14 *Stscr* values are >2.10. Clearly the phylogenetically inferred structure of RNase P RNAs are distinct from random.

In contrast only three optimized structures computed by mfold have high *Stscr* values >2.10. Only for *Thermus aquaticus* RNase P RNA *Stscr* computed from the predicted optimized structure is greater than that from the phylogenetically inferred structure. The *Stscr* scores of the 14 optimized structures average to 1.42 ± 1.59. The sample mean of 14 observations of *Stscr* is far less than the average *Stscr* (*Stscr* = 3.73) computed from the well-established structures. Our data clearly indicate that the phylogenetically conserved structures of RNase P RNAs have more ordered conformations than those in the optimized structures. The difference of structural morphology observed between the phylogenetically inferred and optimized structures is statistically significant. Figure 3 shows that *NRd* and *RRd* cluster completes separately for phylogenetic structures and that *Stscr* cluster differently but not completely separate between phylogenetic and lowest free energy structures.

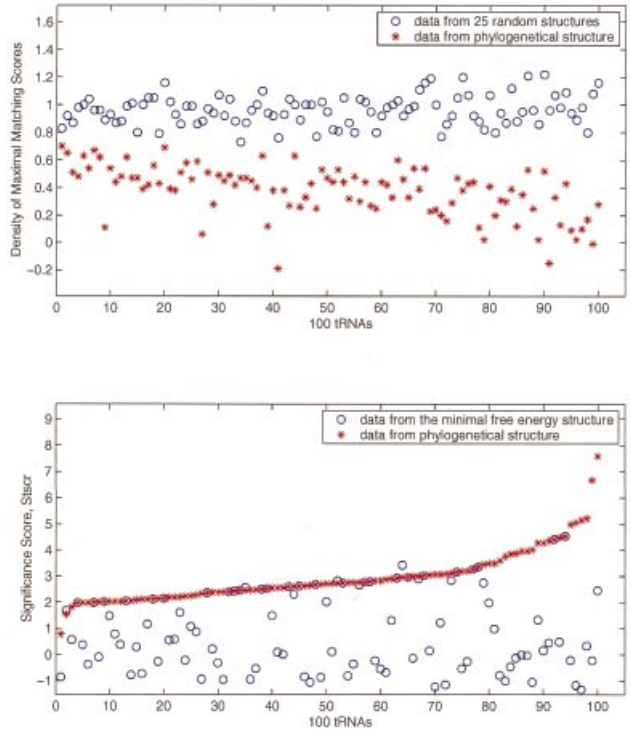

Figure 2. Density values (top) of maximal matching scores, *NRd* and *RRd*, as well as significance scores, *Stscr* (bottom) of the uniqueness of phylogenetically conserved secondary structures and the lowest free energy structures computed from 100 tRNAs. The score *NRd* was computed from structure comparisons between the conserved tRNA secondary structure and 300 random structures that were computed from 300 randomly shuffled sequences of the natural tRNA by mfold and Turner energy rules. The score *RRd* was computed from structure comparisons among 25 random structures and the 300 random structures mentioned above. The 25 random structures were computed from the other 25 randomly shuffled sequences of the natural tRNA by mfold and Turner energy rules. The phylogenetically conserved tRNA secondary structures have small *NRd* and large *Stscr* values.

Some other known RNA elements have well ordered conformations

The results detailed in Table 3 show that FSR elements of *Tetrahymena* rRNA intron, HCV IRES, TAR and RRE of HIV-1 have high values of *Stscr* for the phylogenetically inferred structures. However, the significance score, *Stscr* computed from the optimized, lowest free energy structure is less than that computed from the phylogenetically inferred structures. For example, the significance scores *Stscr* computed from the inferred common structure and the optimized structure by mfold are 4.75 and 0.45, respectively, for the RNA sequence of *Tetrahymena* intron. Similar to tRNAs and RNase P RNAs, the phylogenetic structure of *Tetrahymena* intron has a more ordered conformation than the optimized structure. The computationally optimized structure of TAR functional element predicted by mfold is identical to the phylogenetically inferred structure and the optimized RRE structure is close to the common structure inferred by phylogenetic comparisons. Thus, both the inferred conserved structures and the optimized structures of functional TAR and RRE molecules have high *Stscr* values.

Table 2. Significance scores (*Stscr*) of the uniqueness of RNA secondary structures from 14 RNase P RNA molecules

Sequences		RNA Secondary Structures Computed from Natural RNase P RNA Sequences						Randomly Permuted Sequences		
RNase P RNAs	Size (nt)	Optimized Structures Stscr	NR	NRd	Phylogenetic Structures Stscr	NR	NRd	RR	RRd	std
Anacystis nidulans	385	1.59	147.30	0.38	5.35	56.17	0.15	185.71	0.48	24.20
Agobacterium tumefaciens	402	1.24	139.95	0.35	2.65	90.63	0.23	183.59	0.46	35.12
Bacillus brevis	411	0.28	174.17	0.42	2.45	50.37	0.12	190.10	0.46	57.07
Bacillus subtills	401	5.09	24.81	0.06	5.43	14.67	0.04	175.61	0.44	29.64
Borrelia burgdorferi	411	-0.08	159.27	0.39	2.83	49.23	0.12	156.10	0.38	37.75
Bacteroides thetaiotaomicron	361	1.48	126.50	0.35	3.55	55.26	0.15	177.51	0.49	34.42
Chlorobium limicola	387	2.43	82.81	0.21	3.12	60.08	0.16	163.61	0.42	33.21
Cyanophora paradoxa cyanelle	350	1.07	138.98	0.40	3.80	50.06	0.14	174.02	0.50	32.63
Desulfovibrio desulfuricans	360	-0.26	184.76	0.51	5.43	25.52	0.07	177.57	0.49	27.99
Escherichia coli	377	-0.42	202.93	0.54	2.15	118.67	0.31	189.15	0.50	32.70
Porphyra purpurea chloroplast	383	1.42	133.10	0.35	3.29	71.39	0.19	180.06	0.47	33.01
Streptomyces bikiniensis	398	0.65	172.59	0.43	2.85	69.70	0.18	202.89	0.51	46.80
Thermus aquaticus	395	4.23	95.71	0.24	3.46	116.28	0.29	208.41	0.53	26.64
Thermotoga neapolitana	338	1.21	173.12	0.51	5.84	41.55	0.12	207.39	0.61	28.42

DISCUSSION

The results presented in this paper indicate that FSRs have small *NRd* and large *Stscr*. The distinct structural conformations found in the functional RNA sequences are unlikely to occur by chance. Our data strongly support the hypothesis that RNA molecules, such as tRNAs and RNase P mRNAs, possess well ordered conformations that play a crucial role in their biological functions.

The sequences of FSRs are evolutionary products that have survived because they execute the biological function quite efficiently. It is reasonable to hypothesize that the thermodynamic stability contributed by base pair stacking constrains the extent of possible ordered conformations in RNA evolution. It has been suggested that the evolved conformations of functional RNAs are significantly more ordered than conformations of randomly shuffled sequences (10). Our computational experiment suggests that the sequences of tRNAs, RNase P mRNAs, *Tetrahymena* group I intron and the *cis*-acting elements TAR, RRE and IRES of HCV are optimized with respect to their conformational properties. The sequences of these FSRs have specific structural morphologies to adapt to their particular functions. Multistem junctions are a common morphology in the functional structures of tRNAs, RNase P RNAs, group I intron, IRES of HCV and RRE of HIV, and are also abundant in ribosomal RNAs and small ribonucleoprotein RNAs.

Complete understanding of the role of structured RNAs requires knowledge at the three-dimensional, atomic level. However, we know little about the atomic-level details except what is available for a few tRNAs, the structure core of *Tetrahymena* group I intron and a few RNA fragments. Tinoco and Bustamante (14) proposed that RNA folding is hierarchical and sequential and RNA secondary structure often determines tertiary structure. This implies that correct prediction of RNA secondary structure is a key step for predicting RNA structure from sequence. RNA secondary structures are currently predicted by phylogenetic comparisons and free energy minimization. In calculation of *NRd* and *Stscr* of the test RNAs we use the RNA structures inferred by both methods. Our results show that almost all RNA structures inferred by phylogenetic comparisons tested in this study possess well ordered conformations that are statistically significant. Their distinct conformations are rare features that are not anticipated in randomly permuted sequences. The predicted minimum free energy structures, however, are typically only moderately well ordered. In some cases of extensively studied structures such as HIV TAR and RRE the optimized structures are statistically well ordered.

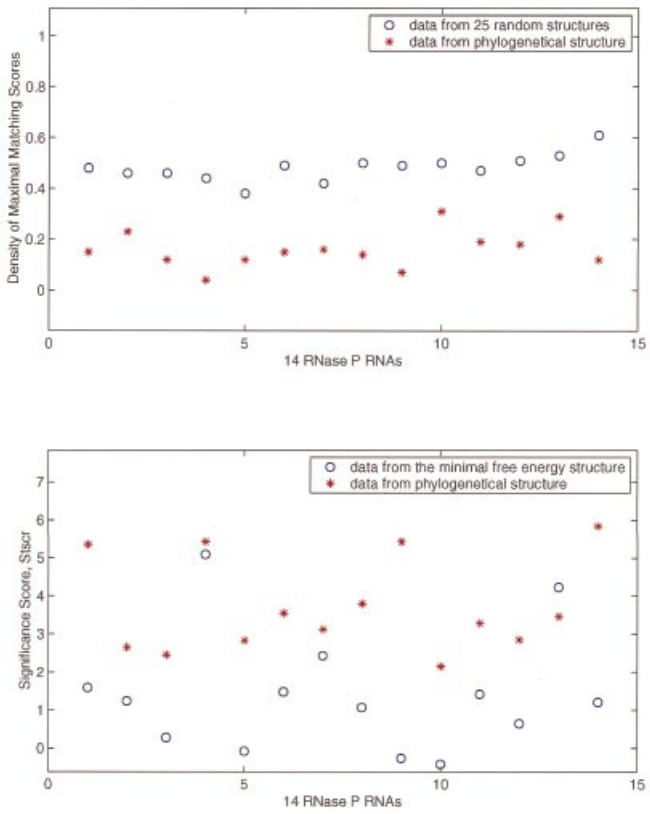

Figure 3. Density values (top) of maximal matching scores, *NRd* and *RRd*, as well as scores *Stscr* (bottom) of the uniqueness of phylogenetically conserved secondary structures and the lowest free energy structures computed from the 14 RNase P RNAs. The phylogenetically conserved secondary structures of RNase P RNAs have small *NRd* and large *Stscr* values. For further details see the legend to Figure 2.

The two quantitative scores of *NRd* and *Stscr* are dependent on score functions associated with the three operations in RNA structure comparison. Two matrices of score functions associated with operations on unpaired bases and paired bases were used in this study. They were derived empirically and arbitrarily. More reasonable score functions need to be developed in future. However, we expect that the significance score *Stscr* of RNA secondary structures may not be sensitive to the score functions used for structure comparison. A test for 34 tRNAs indicates that the difference between *Stscr* scores computed using the two different sets of score functions is not significant.

In the calculation of the measure *RR*, we collect a set of 25 random sequences arbitrarily and compare their folded structures with other 300 random structures. It is also worth choosing those random sequences whose folded energy is less than or close to the lowest free energy computed from the natural sequence in the structure comparison between random versus random sequences. The preliminary results for four tRNAs show that the sample standard deviation computed from random versus random MMS is decreased by the approach. It seems that the approach will slightly improve our method; however, further detailed investigation is needed.

Schultes *et al.* (10) recently proposed three measures to define the stability and uniqueness of RNA secondary structures based on the mean length of stems and total number of base pairs in the computed structure from RNAfold (27) and/or VIENNA (28). The difference between two structure morphologies was not considered throughly in their method. Various similarity measures between two RNA secondary structures have also been discussed by three classes of secondary structure metrics (29). The method used here is related to the 'tree' metrics mentioned in their paper, but does the comparison at a more detailed level. It should be noted that in the work described here only the optimized stable one is selected to represent the random structure and any RNA molecules whose secondary structure was not well established. As a result, conformations used in the structure comparison between natural and randomly shuffled sequences are limited. We know that alternative computed structures with higher free energies can also be computed from the program mfold (16). It is also likely that these alternative structures may be as good as the optimized structure because of uncertainty in the energy parameters and the assumptions used in the mfold algorithm. Nevertheless, our method is the first computational method to examine the detailed difference of structure morphologies throughout between natural RNA molecules and the corresponding randomly shuffled sequences. The results presented in this paper demonstrate that we need a quantitative measure to estimate the uniqueness of the folded RNA structures. The quantitative measure of free energies of RNA folding alone is not enough for us to make a good judgment of a functional RNA structure based on the current energy rules (16).

We previously proposed a computational method using the programs SIGSTB and SEGFOLD (13) to search for unusual folding regions that are thermodynamically more stable than the average of other local segments in the sequence, and more stable than random. The method described here adds the criteria of uniqueness of folding morphologies. We anticipate

Table 3. Significance scores (*Stscr*) of the uniqueness of RNA secondary structures from the functional RNA elements of *T.thermophilia* rRNA intron (T. ther), IRES of HCV, TAR and RRE of HIV-1

		RNA Secondary Structures Computed from								
		Natural RNA Functional Element Sequences						Random Sequences		
Functional	Size	Optimized Structures			Phylogenetic Structures					
RNA Elements	(nt)	Stscr	NR	NRd	Stscr	NR	NRd	RR	RRd	std
TAR of HIV-1	59	2.33	52.01	0.88	2.33	52.01	0.88	69.79	1.18	7.61
RRE of HIV-1	337	2.83	85.35	0.25	5.09	25.01	0.07	160.63	0.48	26.65
IRES of HCV	312	0.28	223.25	0.72	1.80	205.96	0.66	226.38	0.73	11.34
T. ther intron	433	0.45	163.34	0.38	4.75	7.19	0.02	179.84	0.42	36.33

that the combined approach will enhance the ability for data mining of functional RNA elements in mRNAs. However, even with enhanced statistical significance we still need additional biological information to evaluate if a local RNA segment with a well ordered conformation is a functional element. Nevertheless, past experience shows that the approach of computational discovery of structural features is very 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.

ACKNOWLEDGEMENTS

We thank Dr Bruce Shapiro for his insightful suggestions and comments. 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. The program rna_match for structural comparisons can be accessed through the Internet at http://www.csd.uwo.ca/faculty/kzhang/rna/rna_match.html.

REFERENCES

1. Simons,R.W. and Grunberg-Manago,M. (eds) (1998) *RNA Structure and Function*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
2. Bashirullah,A., Cooperstock,R.L. and Lipshitz,H.D. (1998) RNA localization in development. *Annu. Rev. Biochem.*, **67**, 335–394.
3. Gray,N.K. and Wickens,M. (1998) Control of translation initiation in animals. *Annu. Rev. Cell. Dev. Biol.*, **14**, 399–458.
4. Rajagopalan,L.E. and Malter,J.S. (1997) Regulation of eukaryotic messenger RNA turnover. *Prog. Nucleic Acid Res. Mol. Biol.*, **56**, 257–286.
5. Woese,C.R., Winker,S. and Gutell,R.R. (1990) Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proc. Natl Acad. Sci. USA*, **87**, 8467–8471.
6. Gautheret,D., Konings,D. and Gutell,R.R. (1995) G:U base pairing motifs in ribosomal RNA. *RNA*, **1**, 807–814.
7. Elgavish,T., Cannone,J.J., Lee,J.C., Harvey,S.C. and Gutell,R.R. (2001) A:A and A:G base-pairs at the ends of 16 S and 23 S rRNA helices. *J. Mol. Biol.*, **310**, 735–753.
8. Williams,D.J. and Hall,K.B. (2000) Experimental and computational studies of the G(UUCG)C RNA tetra-loop. *J. Mol. Biol.*, **297**, 1045–1061.
9. Draper,D.E. (1996) Strategies for RNA folding. *Trends Biochem. Sci.*, **21**, 145–149.
10. 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.
11. Moore,P.B. (1999) Structural motifs in RNA. *Annu. Rev. Biochem.*, **68**, 287–300.
12. Le,S.-Y., Malim,M.H., Cullen,B.R. and Maizel,J.V.,Jr (1990) A highly conserved RNA folding region coincident with the Rev response element of primate immunodeficiency viruses. *Nucleic Acids Res.*, **18**, 1613–1623.
13. Le,S.-Y., Chen,J.-H. and Maizel,J.V.,Jr (1990) Efficient searches for unusual folding regions in RNA sequences. In Sarma,R.H. and Sarma,M.H. (eds), *Structure & Methods: Human Genome Initiative and DNA Recombination*. Adenine Press, Schenectady, NY, Vol. I, pp. 127–136.
14. Tinoco,I.,Jr and Bustamante,C. (1999) How RNA folds. *J. Mol. Biol.*, **293**, 271–281.
15. Knuth,D. (1973) *The Art of Computer Programming*. Addison-Wesley, Reading, MA, Vol. 3, p. 237.
16. 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.
17. Zuker,M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.
18. Zhang,K. (1998). Computing similarity between RNA secondary structures. *Proceedings of IEEE International Joint Symposia on Intelligence and Systems*. IEEE Press, USA, pp. 126–132.
19. Zhang,K., Wang,L. and Ma,B. (1999) Computing similarity between RNA structures. *Proceedings of the Tenth Symposium on Combinatorial Pattern Matching. Springer-Verlag's Lecture Notes in Computer Science 1645*. Springer-Verlag, Heidelberg, Germany, pp. 281–293.
20. Collins,G., Le,S.-Y. and Zhang,K. (2001) A new algorithm for computing similarity between RNA structures. *Information Sciences*, **139**, 59–77.
21. Eddy,S.R. and Durbin,R. (1994) RNA sequence analysis using covariance models. *Nucleic Acids Res.*, **22**, 2079–2088.
22. Sprinzl,M., Brown,H.M., Ioudovitch,A. and Steinberg,S. (1998) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, **26**, 148–153.
23. Brown,J.W. (1999) The ribonuclease P database. *Nucleic Acids Res.*, **27**, 314.
24. Michel,F. and Westhof,E. (1990) Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.*, **216**, 585–610.
25. Muesing,M.A., Smith,D.H. and Capon,D.J. (1987) Regulation of mRNA accumulation by a human immunodeficiency virus trans-activator protein. *Cell*, **48**, 691–701.
26. Honda,M., Beard,M.R., Ping,L.H. and Lemon,S.M. (1999) A phylogenetically conserved stem–loop structure at the 5′ border of the internal ribosome entry site of hepatitis C virus is required for cap-independent viral translation. *J. Virol.*, **73**, 1165–1174.
27. Zuker,M. and Steigler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–149.
28. 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.
29. Moulton,V., Zuker,M., Steel,M., Pointon,R. and Penny,D. (2000) Metrics on RNA secondary structures. *J. Comput. Biol.*, **7**, 277–292.