J. Biochem. 2009;146(2)251–261
doi:10.1093/jb/mvp064

A Novel Virtual Spectrometry: Visualized Regulatory Motifs on *ADM*, *rPolβ* and *CD83* mRNAs in Human-friendly Manners

Shingo Nakamura*

Frontier Research Laboratories, Takeda Pharmaceutical Company Limited, 10 Wadai, Tsukuba, Ibaraki 300-4293, Japan

Received March 23, 2009; accepted April 6, 2009; published online April 22, 2009

Recently, riboswitches and other structures discovered on mRNAs have been reported as examples of functional RNA structures, motifs. Such motifs were shown to be present as single-form valid structures but they are obscured among other less-valid structures. Here, I present a novel, practical virtual spectrometry (the *GenoPoemics*^{TM} Spectrometry) visualizing motifs on mRNA strands as spectra at-a-glance. Every motif along with validity of their existences could be observed on the spectra in human-friendly manners, and whole structures of mRNAs could be overviewed. Therefore, the spectra helped distinguish valid and less valid motifs. The spectrometry was applied to variety of mRNAs such as *ADM*, *rPolβ* and *CD83* to identify structures of high validity on them, previously reported functional motifs were successfully revealed. These findings indicate that the structures of mRNAs that may be folded into multiple forms can be further discussed quantitatively based on the visual spectra to discover functional RNA motifs.

Key words: mRNA structure, Maxwell-Boltzmann distribution, virtual spectrometry motifs, regulatory motifs, protein binding motifs.

RNA structural biology has developed in a parallel manner, with functional RNAs seeming to also be folded into single forms, as observed for proteins and demonstrated by spectrometric data (here, 'single' means a unique structure for a single sequence). However, the reality is that RNA strands are very rarely folded into only a single form (*1*, *2*). If there is a single optimal form along with multiple sub-optimal forms for a given sequence and the $\Delta\Delta G$ values between the forms are fairly small, the sequence will be folded into multiple forms. In other words, RNA strands having above profiles will be folded into multiple forms, if the time periods for the strands folding under given conditions are very short. Because of that, even short RNA strands did not often seem to be folded into single forms at bench top experiments (regular time periods for folding: seconds to hours), on the other hands, the strands were crystallized and exhibited single 3D structures with non-canonical base interactions (the RNA strands were given enough time for folding. Regular time periods: days to months.). This concept had never been taken into consideration previously simply because RNA sequences characterized by this pattern were presumed to be folded into 'junk' forms without any functionality and were not investigated further. One of the best examples of this was mRNA, which was expected not to be folded into single forms, and thus, was not supposed to be functional.

Recently, structured mRNA (*3–6*) and riboswitches (*7*, *8*) have been reported to control expression of certain genes with or without the assistance of protein. miRNA binding sites on 3′UTR mRNA are likewise likely to be structured (*9–12*). Further, RNA-binding proteins recognize specific mRNA structures, and importantly, they themselves are also recognized by the mRNAs. Generally speaking, structures on the coding sequence (CDS) by causing frameshifts and/or the other unpredictable effects (*13*). Although most of the structures on mRNAs can be classified as 'junk' and are not stably formed, some specific, important local structures are stably formed and are definitely hidden in the mRNAs.

The '*stability*' of motifs has been used to evaluate their importance, and the thermodynamic stability, $\Delta G_{\mathrm{motif}}$, can be calculated with thermodynamic parameters (*14*) by assuming the formation of single structured molecules of the constituent nucleotide sequences of the motifs. However, the existence of the secondary structures is based on both the constituent nucleotide sequence and the neighbouring sequence. Secondary structures, even those with high stability, often cannot be maintained in other regions with different neighboring sequences; that is, particular sequences can be folded into a single form only if the neighbouring sequences cannot interact more tightly with even a part of the constituent sequence. Therefore, the formation of motifs depends not only on the constituent sequences, but also on the neighbouring sequences. Obviously, even motifs showing high 'stability' cannot be biologically important if they do not exist!

MATERIALS AND METHODS

Except for the secondary structure prediction engines, all programs used in the *GP spectrometer* (*gp.1.14.pl or autosampler.5.14.pl.*) along with *GenoPoemics*^{TM} *Viewer++* (GPV++.jar) will be available for download through our web site soon, and is available on a request

*To whom correspondence should be addressed.
Tel: +81-29-864-5034, Fax: +81-29-864-5000,
E-mail: nakamura_shingo@takeda.co.jp

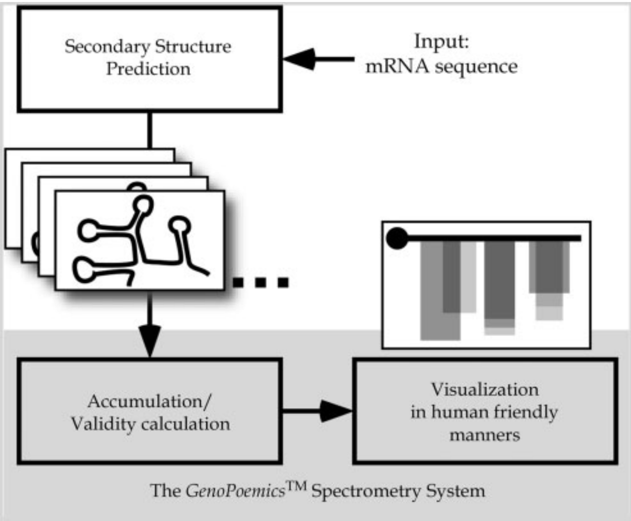
Secondary Structure Prediction
Input: mRNA sequence
Accumulation/ Validity calculation
Visualization in human friendly manners
The GenoPoemics™ Spectrometry System

Fig. 1. **A flow chart of the spectrometry system.** Today's secondary structure prediction software produced many structural candidates including optimal structures for given input RNA sequences. It was impossible to review all the structure candidates, so that some other way to summarize the information has been required. The spectrometry system accumulated the information and visualized local motifs found in the candidates to produce human friendly spectra describing not only variety of structures but also their validities.

basis at present (free for academia). The programs with '.pl' extensions were required for preparing input files (.fld files) for *GenoPoemics*™*Viewer++*, which generated spectra from the input files. The programs and the viewer required a properly configured *Perl 5.8* interpreter and a *Java SE 6.0* environment, respectively. Additionally, installation of *The GCG*® *Wisconsin Package* licensed properly on your server was essential. The mRNA sequence data (NM_001124, U38801 and NM_04233) were all obtained through the NCBI website.

The prediction step and the accumulation step of the prediction results were implemented by server-side programs of the *s*pectrometer system, which were designed to accept *any* secondary structure prediction software as engines and, later, the validity calculation was achieved by the interactive part of the system, the viewer (Fig. 1). It had been reported that *mfold* was not capable of producing all important low free energy structures for given sequences (*15*). However, *mfold* was the *de facto* standard prediction software especially for experimental biologists and widely available as a part of the *GCG package*, so consequently, *mfold* was chosen as the engine for the trials in this article. For detailed information about the software (including installation directions), instructions will be supplied with Supplementary article of this article.

Calculation of the Validity-Static, VS, Based on the Maxwell–Boltzmann Statistics—Even for a single, specific sequence, a number of optimal and sub-optimal structural forms were provided by the secondary prediction software (Figs 1 and 2). The sum of the population of forms in which the motifs were found would be VS. Since spectroscopic methods could not be applied to the study of mRNA, secondary structure prediction software is used to quantitatively compute the validity values. The prediction software could not be directly applied to full-length mRNA because the thermodynamic parameters used in the prediction software were extracted from experiments carried out for small RNA sequences. Based on my experience with *mfold*, reasonable results were obtained for moderate-length RNA sequences of up to about 200 bases. Therefore, for the analysis of full-length mRNA in the *GP spectrometer*, the window for analysis by the prediction software was set to 200 bases in order to gather optimal and sub-optimal predicted secondary structure forms (candidates) in the region (*16*). I assumed that each set of forms could be recognized as a Boltzmann ensemble so that Maxwell–Boltzmann statistics, which were a fundamental thermodynamic concept that could be used to evaluate populations of each structures using ΔG values at equilibrium, could be applied to the resulting output to identify the predicted forms, with each population (P) of candidates calculated as shown below (Equations 1 and 2). Set types of motifs and appropriate parametrical conditions were determined to describe what kinds of motifs were analysed. In this article, parameters were set for single-turn stem loops (simply called 'stem loops' in this article: $\Delta G_{\text{motif}} < 0$ kcal/mol; paired bases >8 bases; mismatch bases < 30 bases; see Fig. 2A for details. The parameters can be fully omitted but it is strongly recommended to set them for reducing calculation time of the motif extraction steps.) and used to screen every form. As a result, the VS$(x)_i$, the sum of the population ratios of candidates among which the motif x was found at the analysis of windows *i* (see Equations 1a, 1b and 2; see also Fig. 2B–D for summary). Additionally, Equation 1b was also set to ignore 'artefact' structures. If the most 3′ side or 5′ side structures found in each windowed sequence were located within 66 bases [the number was given as variable, Exclusion number (Ex), in the system] from both terminals of the window, the motifs were ignored as being artifacts (unless otherwise noted, Equation 1a was used in this article).

$$VS(x)_i[\%] = \frac{100}{P_{i,\text{total}}} \sum_{j=1}^{m} \begin{cases} +P_{i,j} & \text{If the motif } x \text{ is found in form } j \\ +0 & \text{If the motif } x \text{ is NOT found in form } j \end{cases} \tag{1a}$$

$$VS(x)_i[\%] = \frac{100}{P_{i,\text{total}}} \sum_{j=1}^{m} \begin{cases} +P_{i,j} & \text{If the motif } x \text{ is found in form } j \text{ and the motif is NOT the most terminal motif in window } i \\ +0 & \text{If the motif } x \text{ is NOT found in form } j \text{ or the motif is the most terminal motif in window } i \end{cases} \tag{1b}$$

where *m* is the number of the candidates predicted for window *i*.

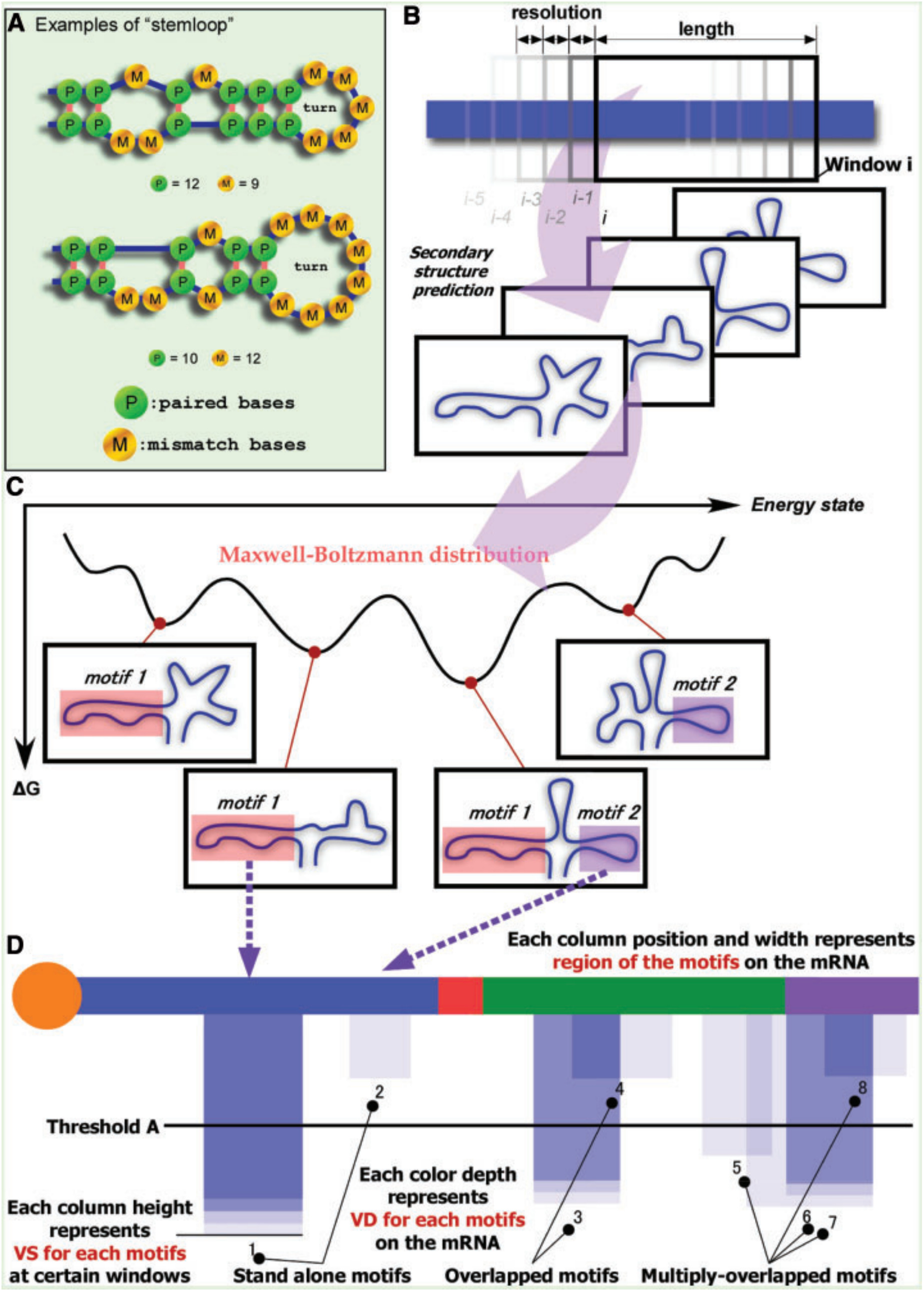
A Examples of "stemloop"
P
M
turn
P = 12
M = 9
P = 10
M = 12
P :paired bases
M :mismatch bases
B
resolution
length
Window i
i-5
i-4
i-3
i-2
i-1
i
Secondary structure prediction
C
Energy state
Maxwell-Boltzmann distribution
ΔG
motif 1
motif 2
D
Each column position and width represents region of the motifs on the mRNA
Threshold A
1
2
3
4
5
6
7
8
Each column height represents VS for each motifs at certain windows
Each color depth represents VD for each motifs on the mRNA
Stand alone motifs
Overlapped motifs
Multiply-overlapped motifs

Fig. 2. **Calculation of VS and VD.** (A) Two examples of typical motifs used in the *GenoPoemics*$^{\mathrm{TM}}$ system, single-turn stemloops. If given conditions matched, motifs were identified and extracted for the VS(x)i calculation. The conditions were described as type of motifs, required $\Delta G_{\mathrm{motif}}$ values, ranges of paired bases and mismatch bases. Typical conditions: single-turn stemloop; $\Delta G_{\mathrm{motif}} < 0.0$ [kcal/mom]; paired bases > 8; mismatches bases < 30. (B) A window of a set length produced 'windowed' sequence fragments of a set length. The sequence fragment was processed with secondary structure prediction software to produce a set of predicted secondary structures. (C) The set of predicted secondary structures (forms) with corresponding ΔG for a windowed sequence were obtained by secondary structure prediction software. The population of each form could be calculated by Maxwell–Boltzmann statistics as if the forms were at the equilibrium state. VS(x)i was calculated with Eqs 1–3 for each motif extracted. Multiple motifs could be identified and extracted in a single window. (D) A window was moved iteratively at a set resolution from 5′ to 3′ along the transcript to produce the 'windowed' sequence fragments (as shown A). Extracted motifs corresponding to each windowed sequence were visualized as columns at their absolute position on the mRNA. Pale blue to black columns represented the position of regions containing stem loops, with the width representing the length of the motif and height representing the VS of the stemloops in each window. A motif identified in different analysis windows as having the same length and absolute position was depicted as a set of fully overlapping columns (*1*). On the other hand, multiple motifs (1 and 2) were often found in a single window. Motifs that were sufficiently close in proximity were represented by partially overlapping columns (3 + 4 and 5 + 6 + 7 + 8). Given the threshold A for VS, column 4 would be ignored and column 3 would be recognized as a stand-alone motif like column 1 (recognized with or without the threshold condition). However, under the same conditions, none of columns 5, 6, 7 or 8 would be recognized as stand-alone motifs. The colour intensity of columns for stand-alone motifs represented VD.

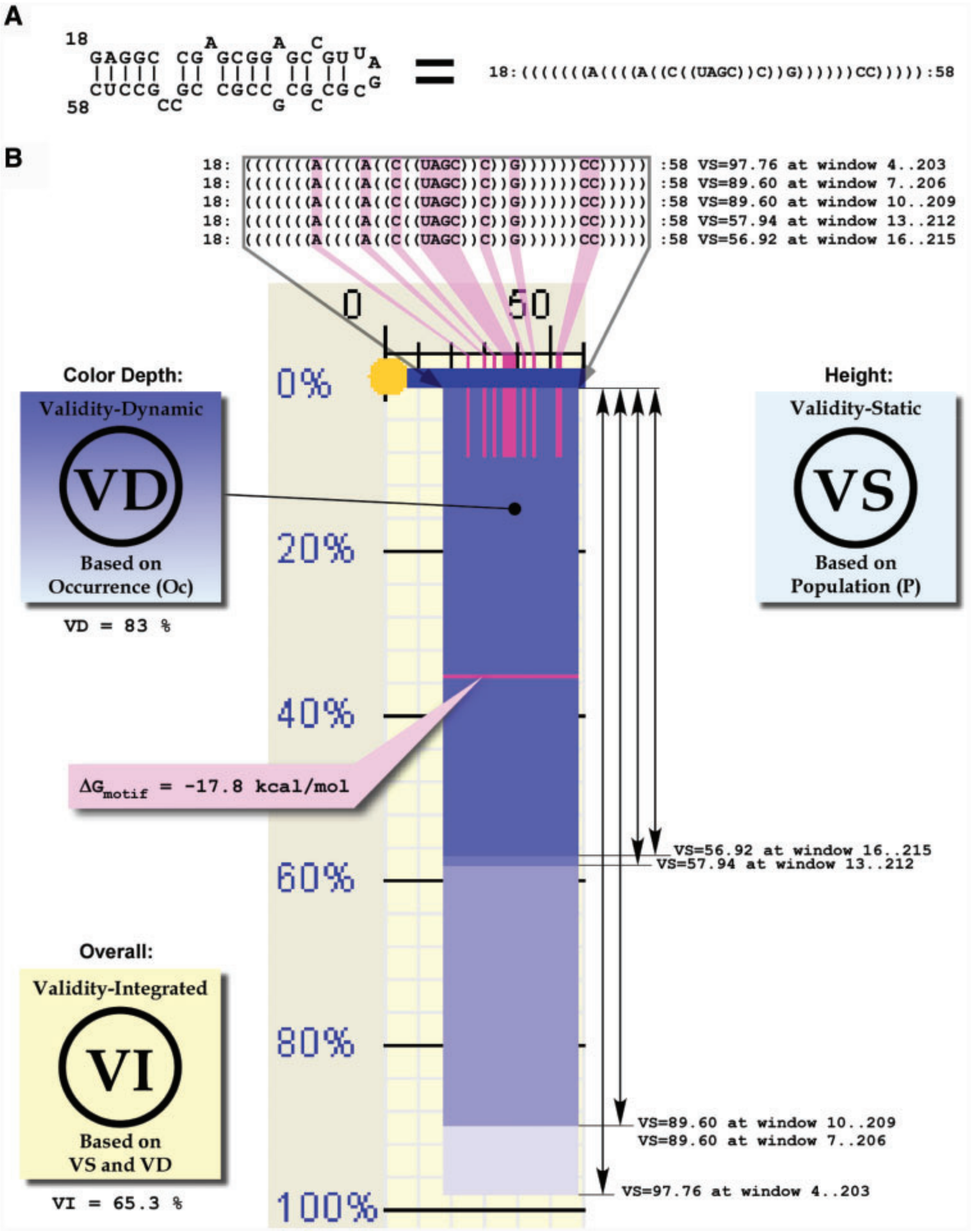
A
18
GAGGC CG A GCGG A GC C GU U A
CUCCG GC CGCC CG CGC G
58 CC G C
18:(((((((A((((A((C((UAGC))C))G))))))CC))))):58
B
18: (((((((A((((A((C((UAGC))C))G))))))CC))))) :58 VS=97.76 at window 4..203
18: (((((((A((((A((C((UAGC))C))G))))))CC))))) :58 VS=89.60 at window 7..206
18: (((((((A((((A((C((UAGC))C))G))))))CC))))) :58 VS=89.60 at window 10..209
18: (((((((A((((A((C((UAGC))C))G))))))CC))))) :58 VS=57.94 at window 13..212
18: (((((((A((((A((C((UAGC))C))G))))))CC))))) :58 VS=56.92 at window 16..215
0
50
0%
20%
40%
60%
80%
100%
Color Depth:
Validity-Dynamic
VD
Based on Occurrence (Oc)
VD = 83 %
Height:
Validity-Static
VS
Based on Population (P)
ΔGmotif = -17.8 kcal/mol
VS=56.92 at window 16..215
VS=57.94 at window 13..212
Overall:
Validity-Integrated
VI
Based on VS and VD
VI = 65.3 %
VS=89.60 at window 10..209
VS=89.60 at window 7..206
VS=97.76 at window 4..203

Fig. 3. **An example of a stemloop and columns on spectra.** (A) Stemloop structures could be described as strings. A stemloop found on 18–58 of an mRNA was drawn as a figure and a string. In this string, each pair of parentheses [‘(’ and ‘)’] represented each pair of base pairing of the stem, and characters represented free bases such as bases of mismatches, bulges or loops. Therefore, successive characters in a pair of parentheses (‘UAGC’, in this case) represented loop region. (B) Columns on spectra. Position, height and colour intensity represented corresponding absolute position, VS and VD of motifs, respectively. Each vertical bar at the base line represented free bases of motifs. $\Delta G_{\mathrm{motif}}$ was also shown as a horizontal bar.

$$P_{i,\mathrm{total}} = \sum_{j=1}^{m} P_{i,j} \tag{2}$$

$$P_{i,j} = \frac{\exp\left(\frac{-\Delta G_{i,j}}{RT}\right)}{\exp\left(\frac{-\Delta G_{i,\min}}{RT}\right)} \tag{3}$$

where $\Delta G_{\mathrm{i,min}}$ is the lowest free energy amongst the candidates predicted for window i.

Calculation of the Validity-Dynamic, VD and the Validity-Integrated, VI—Even if the sequences contained motif as sequences, the structural motifs were not always appeared in the predicted structures of the sequences. VD represented how often motifs maintained the same forms, even under the possible different sequence conditions (Fig. 3). To analyse full-length mRNA, the sequence window was applied to the entire mRNA from the 5′ to the 3′ terminus. Additionally, to collect VS data for every motif, surrounded by every possible sequence, the 200-base window was set to overlap, moving along the strand 3 bases (set as the resolution) at a time to match the length of a codon. The windows thus covered every possible 200-base sequence, encompassing the motifs and surrounding mRNA. VD was then calculated according to Equation 4 to show how often the specific motif is found in the series of windows. As VD was calculated from occurrence numbers (defined as Oc), W(*i, x*), a weighting function, had been proposed and was tentatively set to 1 to simplify the model (see Equations 4–6; see also Fig. 2 for summary). Therefore, in this article, $\mathrm{Oc}_{\mathrm{total},x}$ equals to n, number of windowed sequences that include the constituent sequence of the motif *x*.

$$VD(x)_i[\%] = \frac{100}{Oc_{\mathrm{total},x}} \begin{cases} +Oc(x)_i & \text{If the motif } x \text{ is found} \\ & \quad \text{in window } i \\ +0 & \text{If the motif } x \text{ is NOT} \\ & \text{found in window } i \end{cases} \tag{4}$$

$$Oc_{\mathrm{total},x} = \sum_{i=1}^{n} Oc(x)_i \tag{5}$$

$$Oc(x)_i = \begin{cases} +W_{(i,x)} & \text{If the motif } x \text{ is found} \\ & \quad \text{in window } i \\ +0 & \text{If the motif } x \text{ is NOT found} \\ & \quad \text{in window } i \end{cases} \tag{6}$$

where *n* is number of windowed sequences that include the constituent sequence of the motif *x*.

Finally, VI, validity-integrated which represents the comprehensive validity of motifs, was calculated as follows (Equation 7).

$$VI(x)[\%] = \frac{1}{100} \sum_{i=1}^{n} (VS(x)_i \times VD(x)_i) \tag{7}$$

Visualization—Motifs are influenced not only by the surrounding sequences, which are taken into consideration in the VD calculation, but also by surrounding motifs. Therefore, visualizing the output as 'spectra' arranged by position along the sequence allowed viewing of all motifs at a glance. Among the spectra, blue columns indicated alignment with the mRNA sequences, with the start of the column corresponding to the motif start position, and the width and height representing the motif length and VS, respectively (Fig. 3). Columns having greater overlap showed deeper colour intensity—deep colours represented VD. Deeper colour intensity also appeared where columns of different motifs overlaped (Fig. 4A). The spectrum also provided other types of information: $\Delta G_{\mathrm{motif}}$ of each motif and free/pseudo-free base information (Figs 3B and 4B; also see the Supplementary Material). As a result, in the *GenoPoemics*$^{\mathrm{TM}}$ Spectrometry System, three numbers, validity-static, validity-dynamic and validity-integrated, of mRNAs were calculated and evaluated, and finally the system produced visualized spectra-at-a-glance, which was comprehensive accumulation of not only motifs found on mRNAs but also their validities. The motifs that were not qualified for threshold values mainly set for the validity values were omitted and did not show up on the spectra so that properly configured threshold values gave simple spectra having possible important structures. The proper threshold values could be determined along with experimental results. For example, if threshold VI was given as 90%, the motifs having less VI values were all ignored. The values were configured interactively on the viewer. Such threshold values were to be set for VS, VI, VD and $\Delta G_{\mathrm{motif}}$ in this article.

RESULTS

Virtual Spectrometer—A concept, '*validity*', was first introduced along with '*stability*' to evaluate the presence of motifs (tiny structures, stem loops for example). Validity of certain motifs is the probability of their existence under biological conditions. It is generally true that more stable structures tend to be more valid based on their existence, even though their existence depends on the neighbouring sequence environment: again, the stable structures exist only if the sequences of the structures do not interact with adjacent sequences such that they are not folded into different forms. Structures with high validity may have biological significance. Certain existence of RNA structures was often confirmed by X-ray crystal analyses, but these analyses cannot be applied to long RNAs. This is again because long RNAs cannot be folded into single forms and, basically, long RNAs cannot be crystallized easily. Thus, a possible realistic method to handle these structures is using *in silico* prediction software (*17*) (Fig. 1). Validity is calculated from the output of the *de facto* standard software program, *mfold* (*18*, *19*), its improved version, *UNAFold* (*20*) and alternative software, *Sfold* (*21*), Vienna RNA package (*16*), *vsfold* (*22*) and others. These programs can identify optimal secondary structure forms with or without suboptimal structures for given sequences and calculate the corresponding stability in terms of free energy (ΔG). However, these programs were developed using experimental data from short RNA sequences (*14*) and in practical application to long RNAs, generally too many proposed structures are produced at once. Therefore, a new method was needed to visualize comprehensively and human-friendly how mRNAs are folded in media for further investigation. In other words, a new *virtual* spectrometer, based on the *in silico* secondary structure prediction software, was needed to observe not just structures of mRNAs but also their validity.

The *true* validity of motifs was divided into the static and dynamic validity of motifs, with the integrated validity being the *true* validity. The validity-static (defined as VS) described the validity of motifs in static sequence conditions, where motifs were distributed as a number of optimal and sub-optimal structural forms for a single, specific sequence, and the sum of the population of forms in which the motifs were found would be VS (Fig. 2B and C). On the other hand, validity-dynamic (defined as VD) described how often motifs maintain the same forms, even under different possible sequence conditions (Fig. 2D). Therefore, VD described how motifs maintain the VS, even under different possible static conditions so that properly accumulated VS information under different possible sequence conditions led to the definition of validity-integrated (defined as VI), which was considered to incorporate both VS and VD. The VI represented the *true* validity of motifs. These numbers, VS, VD and VI, were quantities and represented as

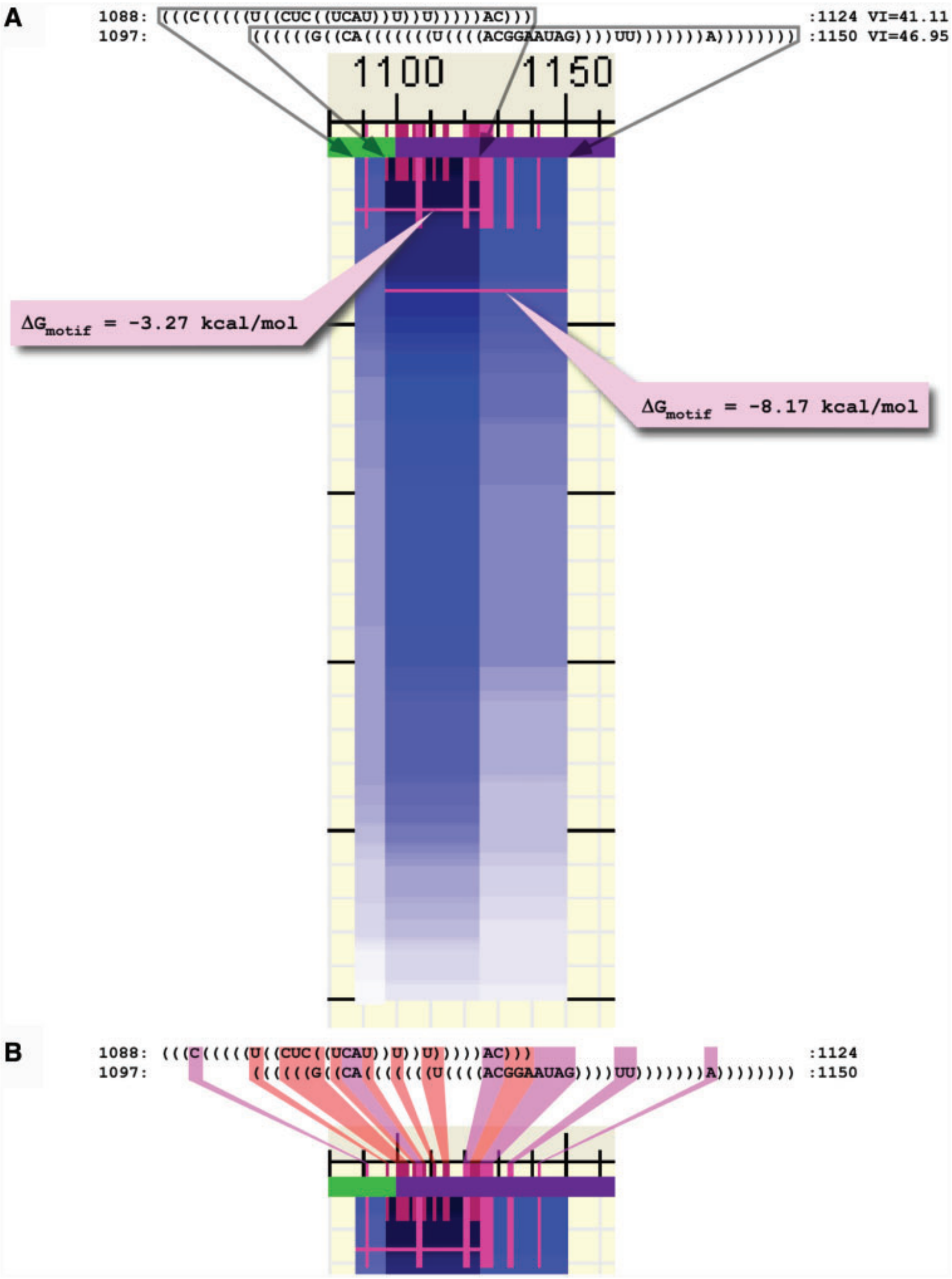
A
1088: (((C(((((U((CUC((UCAU))U))U)))))AC))) :1124 VI=41.11
1097: ((((((G((CA(((((((U((((ACGGAAUAG))))UU)))))))A)))))))) :1150 VI=46.95
1100
1150
ΔGmotif = -3.27 kcal/mol
ΔGmotif = -8.17 kcal/mol
B
1088: (((C(((((U((CUC((UCAU))U))U)))))AC))) :1124
1097: ((((((G((CA(((((((U((((ACGGAAUAG))))UU)))))))A)))))))) :1150

Fig. 4. **An example of multiple-overlapping columns.** (A) Obviously, the darker column (1097–1124) was a column originated from two different motifs; therefore, it did *NOT* represent any single motif. (B) Free bases were represented as long vertical bars along the sequence; these bases did not take part in stems of any motif represented by the columns. Pseudo-free bases were represented as short vertical bars along the sequence; these bases did not take part in stems of some motifs represented by the columns but participated in the stems of other motifs. The free/pseudo-free bases were to be used to distinguish among columns originating from different motifs. In other words, if the pseudo-free bases were observed in overlapping columns, the columns originated from at least two different motifs.

valued numbers in the practical virtual spectrometry system, the *GenoPoemics*^{TM} Spectrometry System (the *GP spectrometer*). In the *GP spectrometer*, not just optimal structures, but also suboptimal structures, were taken into consideration. Finally, the *GP spectrometer* visualized motifs found in the structures as human-friendly manners (spectra) that were not only plausible based on stability, but which also had a sufficiently high possibility of existence (*validity*).

Application of the analysis method to *ADM*, *rPolβ*, and *CD83* mRNA produced structural spectra (Figs 5–7). In mRNA spectra, sequences were represented schematically across the top as a bar with the following components represented by colors and symbols: CAP, orange

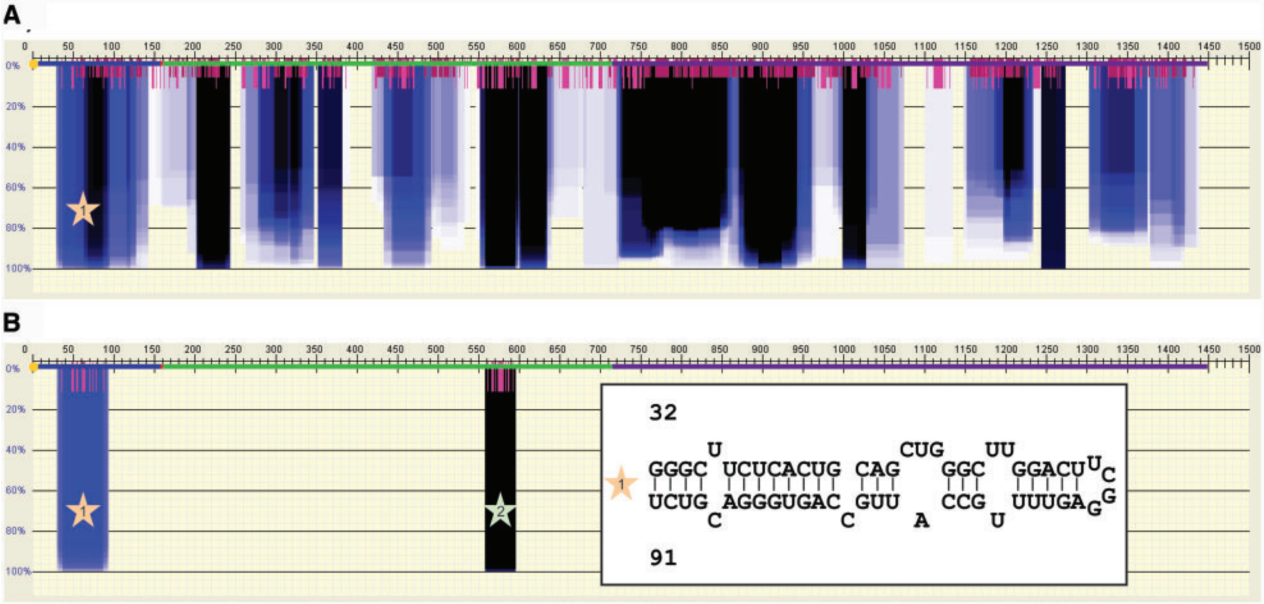

Fig. 5. **Spectra of *ADM* mRNA.** (A) The intact spectrum. (B) Threshold VI = 90% was applied.

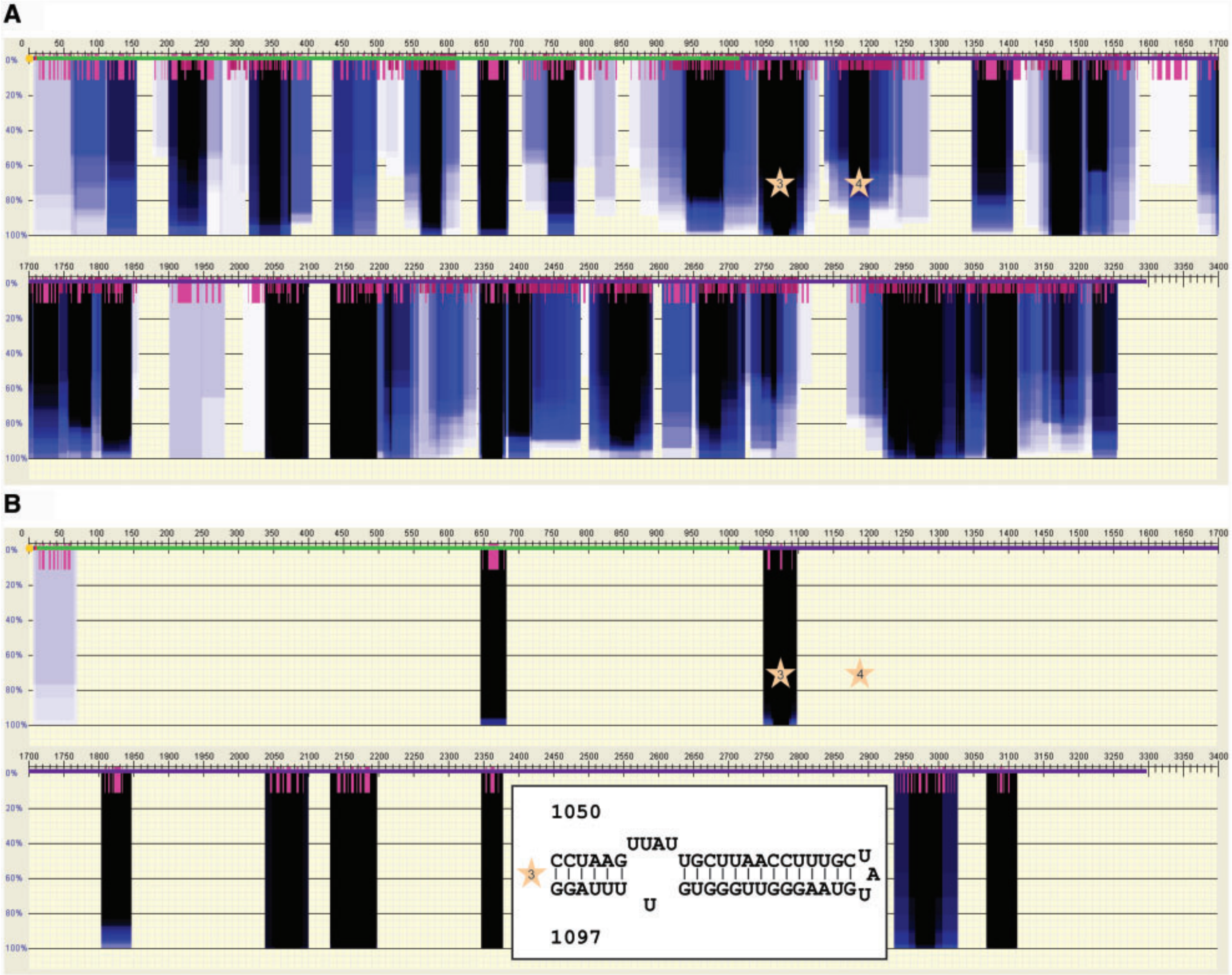

Fig. 6. **Spectra of *rPolβ* mRNA.** (A) The intact spectrum. (B) Threshold VI = 90% was applied.

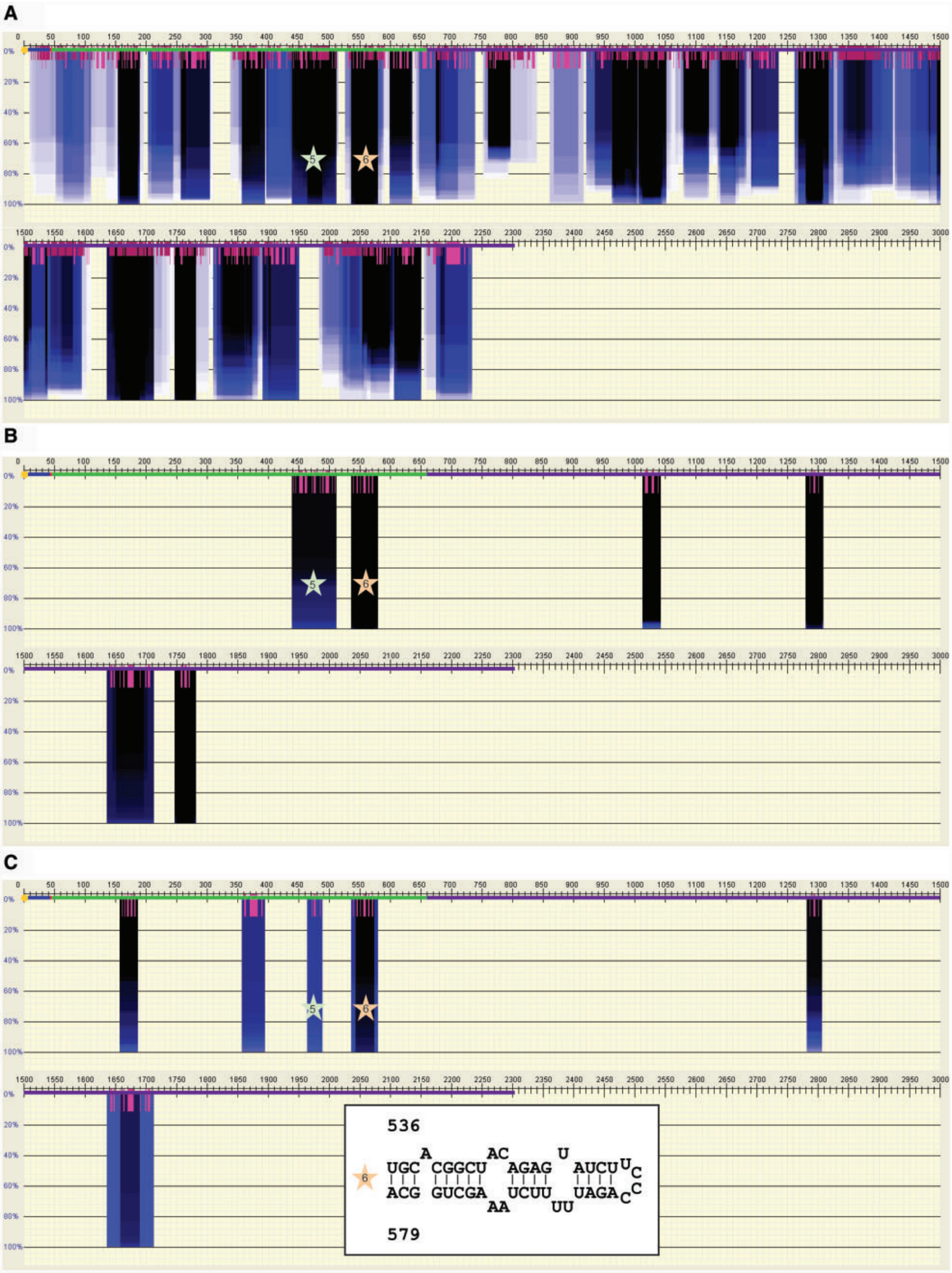

Fig. 7. **Spectra of *CD83* mRNA.** (A) The intact spectrum. (B) Threshold VI = 90% was applied. (C) Threshold VI = 50% and 5′exclusion were applied.

circle at the 5′-end; 5′-UTR, blue bar; AUG start codon, red bar; CDS, green bar; and 3′-UTR, purple bar. Visualizing the output as 'spectra' arranged by position along the sequence allowed viewing of all motifs at a glance. Among the spectra, blue columns indicated alignment with the mRNA sequences, with the start of the column corresponding to the motif start position, and the width and height representing the motif length and VS, respectively. Columns having greater overlap showed deeper color intensity—deep colours represented VD. Deeper color intensity also appeared where columns of different motifs overlaped. Appropriate thresholds for VS, VD, VI and $\Delta G_{\mathrm{motif}}$ were applied to identify meaningful motifs (indicated with stars in these spectra) that had previously been reported by conventional research methods.

Spectra: 5′-UTR of ADM mRNA—Brenet *et al.* (*23*) reported a stem loop structure found on the 5′-UTR of human *adrenomedullin (AM)* mRNA. AM is one of the multifunctional regulatory peptides that is well known for its important angiogenic and mitogenic properties. As is often described, the structures on 5′-UTR were expected to exert a great influence on the translation levels of the proteins encoded thereafter (*3*, *4*, *6*), and experimental results clearly showed that the stem loop caused post-transcriptional regulation of *AM* gene expression. Analysis of the mRNA sequence (NM_001124) by the *GP spectrometer* (win = 200, res = 3; shown in Fig. 5A) and following application of a threshold VI (90%) to exclude unmatched columns, identified two columns (Fig. 5B). The stemloop 1, represented by the first column (bases 32–91), matched the stem loop reported. Application of an additional threshold (the threshold for $\Delta G_{\mathrm{motif}} = -10$ kcal/mol) resulted in only stem loop 1 ($\Delta G_{\mathrm{motif}} = -23.33$ kcal/mol) remaining unaffected on the spectrum (stem loop 2: $\Delta G_{\mathrm{motif}} = -9.72$ kcal/mol).

Spectra: 3′-UTR of rPolβ mRNA—Sarnowska *et al.* (*24*) reported a stem loop structure found on the 3′-UTR of *rat DNA polymerase β* (*rPolβ*) mRNA. The rPolβ is an essential enzyme for base excision repair, and aberrant expression of the protein leads to genetic instability and carcinogenesis. The structures of the region act as post-transcriptional regulatory elements and interact with the Hax-1 protein, an anti-apoptotic, cytoskeleton-related protein, which is known to bind a stem loop structure within the 3′-UTR of *vimentin* mRNA. The experimental results of Sarnowska *et al.* demonstrated stem loops controled of the level of translation. Analysis of the mRNA sequence, U38801, by the *GP spectrometer* (Fig. 6A) showed stem loops M1 (not marked in the spectrum; VI = 4.32%), M2 (stem loop 3: VI = 90.83%), and part of M3 (stem loop 4: VI = 40.21%). Application of threshold VI (90%) excluded unmatched columns, leaving nine column groups (Fig. 6B). The stem loop 3 represented by the column around 1050–1097 was the same stem loop as M2, which is emphasized as a key stem loop.

Spectra: CDS of CD83 mRNA—Prechtel *et al.* (*25*) reported a stem loop structure found on the CDS of *CD83* mRNA. The CD83 protein is the best marker among all the known proteins for the fully mature dendritic cells (DC). Although its exact function remains undetermined, it is thought to play an important role in DC-mediated T-cell immunity. Experimental results clearly showed that the stem loop controlled the level of translation together with HuR protein, a member of a family of human RNA-binding proteins that is related to the *Drosophila* embryonic lethal abnormal vision (ELAV) protein. Analysis of the corresponding mRNA (NM_04233) by the *GP spectrometer* (Fig. 7A) and with application of a threshold VI (90%) to exclude unmatched columns, identified six columns (Fig. 7B). The stem loops represented as the first column (stem loop 5: 439–511) were similar to SL1 and the second column (stem loop 6: 536–579) was the same as a part of SL2. Stem loop 5 was NOT exactly the same as SL1, but the motif occupied the region of SL1 and even exhibited high stability, so that the motif was assumed as an SL1 equivalent in this article. After applying Equation 1b instead of Equation 1a (see MATERIALS AND METHODS section for details) and an appropriate threshold for VI (50%) (Fig. 4C), stem loop 6 was determined to have greater validity than stem loop 5 (only the upper part of stem loop 5 appeared on Fig. 7C.) and that matched the experimental results, in which HuR only interacted with SL2, not SL1.

The spectra produced by the *GP spectrometer* had parallels to spectra produced by spectrometers, and this system of viewing and evaluating the validity along the length of a sequence allowed analysis of overall structures on mRNAs. Although many columns representing motifs were shown on the raw spectra, most motifs had no biological importance, mostly because they were only just deemed as being valid. The previously reported important structures on *ADM*, *rPolβ* and *CD83* mRNA could be observed, particularly following application of appropriate threshold values. According to these three results, highly valid stem loops (shown even threshold of VI = 90%) seem to be functional in the cell. The stability of the motifs was partly evaluated in the VS calculation step, therefore, making the additional evaluation of the values for the motifs along with validity non-essential, but optional. The additional threshold of $\Delta G_{\mathrm{motif}}$ (−10 kcal/mol) for *ADM* mRNA spectrum was suitable because the event involving the motif was dependant on the $\Delta G_{\mathrm{motif}}$ of the motif. This observation corroborates the theory that the occurrence of the event is the result of a competition between the ribosome helicase activity to ravel structures at the region of the mRNA, and stability of the motif to resist this raveling.

Protein binding at sites on *rPolβ* and *CD83* mRNAs could be discussed using these spectra. Since chaperone proteins have not been widely reported for mRNAs to date, most observations were of proteins that bind to RNA structures that were folded prior to the binding events, and it seemed possible to discuss motifs involved in RNA–protein interactions based on only these spectra. Some other 'valid' stem loops had been observed, suggesting that other factors played a role in making the reported motifs unique, such as, an interaction with other valid structures, the general level of complexness of the adjacent area, or biological properties of regions. At the same time, some motifs were shown to be invalid even though they had been claimed valid in the previous report by *in vitro*. Especially, in the case of *rPolβ*, where M1 and M3 were shown to have VI = 4.32% and

VI = 40.21%, respectively. M1 was found to be unstable ($\Delta G_{\mathrm{motif}} = -4.77$ kcal/mol) and seemed not to exist *in situ* therefore it was not important for interaction of the mRNA with Hax-1. It was possible that Hax-1 might bind M2 first, then help in the formation of M3, like RNA chaperones causing M3 itself to exhibit low validity on the spectrum. The system could act as a stimulus for comprehensive discussion on such occurrences.

On the other hand, some false-positive 'artefact' structures, which were produced only as a result of the given conditions of the system, had no biological meaning, and consequently they had been ignored. The application of Equation 1b to negate 5′ terminal artifacts could improve this approach, as demonstrated by the case of *CD83* mRNA. Actually, Ex was variable in GPV, and Equation 1a equals to Equation 1b if zero (default) was given for Ex. What Ex number is suited to the analyses remains uncertain. Various patterns of Ex along with other parameters in future must be subject to further analysis in future.

While a certain number of the other false positive motifs appeared on the spectra only because of imperfections in factors/variables, such as calculation conditions, parameters, and the *GP spectrometer* logics, truly valid motifs that existed on the spectra were likely to actually exist. Appropriate sets of parameters, starting with validity, could be selected in order to identify motifs that were likely to be valid. Having identified potential valid motifs on the spectra, biological experiments could then be focused efficiently to confirm the biological function of these motifs. This virtual spectrometry and the wet bench techniques also raise the possibility that unexpected positive motifs, that is, motifs identified on the spectra but not confirmed with the experiments, may simply have an unknown biological function.

DISCUSSION

The motifs evaluated by the *GP spectrometer* in this report were on 5′-UTR, CDS and 3′-UTR regions of the respective mRNAs, suggesting that the system could be applied to analyze the entire mRNA sequence. Events could be revealed more easily if appropriate threshold numbers were applied to unknown mRNA sequences to spot stable and valid motifs. By extension, other parameters such as window length and resolution could be adjusted parametrically or functionally. Furthermore, not only stem loops but also other motifs of interest, such as gaps between stem loop-like structures ('groin' motif; see the Supplementary Material), could be subjected to the spectral analysis.

The relative merits of alternative prediction software suitable for practical structural study of RNA have been subject to intensive discussion (*15*, *21*, *23*, *26–28*). Thus, the system was designed to be independent from the secondary prediction software. Prediction software produce limited numbers of optimal and suboptimal structure candidates, therefore conditions for the prediction software must be tuned to realize that the sets of candidates can represent Boltzmann ensembles. The *GP* spectrometer was a proposal of new viewpoint for mRNA structures at this time moment: the sensitivity and specificity of the spectrometer system must be verified regularly later along with the various prediction software.

In conclusion, using the *GP* virtual spectrometry system, properties of stem loops, which function in the cell, could be determined based on Maxwell–Boltzmann statistics and visualized. Although the application of this method had not been verified for pre-mRNAs and ncRNAs yet, it was applicable to mRNAs. The visualized quantitative validity values serve as a starting point for the study of long RNA structures that do not assume single, unique structures. The system was knowingly based on simple reliable technologies and produced spectra of raw structures of mRNAs to leave experimental biologists rooms for their own analyses to solve their own scientific issues at their own desks. We do NOT yet comprehend the structures of mRNAs enough to conclude any, so that experimental results along with the raw spectra would help proceed with the structural studies of mRNAs.

SUPPLEMENTARY DATA

Supplementary data are available at *JB* online.

ACKNOWLEDGEMENTS

I would like to thank Prof. Dr Michael Zuker for his helpful discussions. I appreciate the continuous support and encouragement of Drs Nobuhiro Suzuki, Susumu Honda and Shigenori Ohkawa.

CONFLICT OF INTEREST

None declared.

REFERENCES

1. Russell, R. (2008) RNA misfolding and the action of chaperones. *Front Biosci.* **13**, 1–20
2. Herschlag, D. (1995) RNA chaperones and the RNA folding problem. *J. Biol. Chem.* **270**, 20871–20874
3. Kozak, M. (1986) Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. *Proc. Natl Acad. Sci. USA* **83**, 2850–2854
4. Kozak, M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361**, 13–37
5. Pickering, B.M. and Willis, A.E. (2005) The implications of structured 5′ untranslated regions on translation and disease. *Semin. Cell Dev. Biol.* **16**, 39–47
6. Ringnér, M. and Krogh, M. (2005) Folding free energies of 5′-UTRs impact post-transcriptional regulation on a genomic scale in yeast. *PLoS Comput Biol.* **1**, e72
7. Barrick, J.E. and Breaker, R.R. (2007) The power of riboswitches. *Sci. Am.* **296**, 50–57
8. Winkler, W.C. (2005) Riboswitches and the role of noncoding RNAs in bacterial metabolic control. *Curr. Opin. Chem. Biol.* **9**, 594–602
9. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet.* **39**, 1278–1284
10. Rana, T.M. (2007) Illuminating the silence: understanding the structure and function of small RNAs. *Nat. Rev. Mol. Cell Biol.* **8**, 23–36
11. Collins, R.E. and Cheng, X. (2006) Structural and biochemical advances in mammalian RNAi. *J. Cell Biochem.* **99**, 1251–1266

12. Song, J.J., Smith, S.K., Hannon, G.J., and Joshua-Tor, L. (2004) Crystal structure of Argonaute and its implications for RISC slicer activity. *Science* **305**, 1434–1437
13. Alam, S.L., Wills, N.M., Ingram, J.A., Atkins, J.F., and Gesteland, R.F. (1999) Structural studies of the RNA pseudoknot required for readthrough of the gag-termination codon of murine leukemia virus. *J. Mol. Biol.* **288**, 837–852
14. Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911–940
15. Mathews, D.H. (2006) Revolutions in RNA secondary structure prediction. *J. Mol. Biol.* **359**, 526–532
16. Hofacker, I.L., Priwitzer, B., and Stadler, P.F. (2004) Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics* **20**, 186–190
17. Mathews, D.H., Turner, D.H., and Zuker, M. (2007) RNA secondary structure prediction. *Curr. Protoc. Nucleic Acid Chem.* Chapter 11:Unit 11.2.
18. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415
19. Dimitrov, R.A. and Zuker, M. (2004) Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys. J.* **87**, 215–226
20. Markham, N.R. and Zuker, M. (2005) DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.* **33**, W577–W581
21. Ding, Y. and Lawrence, C.E. (2003) A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.* **31**, 7280–7301
22. Dawson, W., Fujiwara, K., Kawai, G., Futamura, Y., and Yamamoto, K. (2006) A method for finding optimal rna secondary structures using a new entropy model (vsfold). *Nucleosides Nucleotides Nucleic Acids* **25**, 171–189
23. Brenet, F., Dussault, N., Delfino, C., Boudouresque, F., Chinot, O., Martin, P.M., and Ouafik, L.H. (2006) Identification of secondary structure in the 5′-untranslated region of the human adrenomedullin mRNA with implications for the regulation of mRNA translation. *Oncogene* **25**, 6510–6519
24. Sarnowska, E., Grzybowska, E.A., Sobczak, K., Konopinski, R., Wilczynska, A., Szwarc, M., Sarnowski, T.J., Krzyzosiak, W.J., and Siedlecki, J.A. (2007) Hairpin structure within the 3′UTR of DNA polymerase beta mRNA acts as a post-transcriptional regulatory element and interacts with Hax-1. *Nucleic Acids Res.* **35**, 5499–5510
25. Prechtel, A.T., Chemnitz, J., Schirmer, S., Ehlers, C., Langbein-Detsch, I., Stülke, J., Dabauvalle, M.C., Kehlenbach, R.H., and Hauber, J. (2006) Expression of CD83 is regulated by HuR via a novel *cis*-active coding region RNA element. *J. Biol. Chem.* **281**, 10912–10925
26. Wuchty, S., Fontana, W., Hofacker, I.L., and Schuster, P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* **49**, 145–165
27. McCaskill, J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29**, 1105–1119
28. Bompfünewerer, A.F., Backofen, R., Bernhart, S.H., Hertel, J., Hofacker, I.L., Stadler, P.F., and Will, S. (2008) Variations on RNA folding and alignment: lessons from Benasque. *J. Math Biol.* **56**, 129–144