© 1995 Oxford University Press
Nucleic Acids Research, 1995, Vol. 23, No. 3 419–426

A genetic algorithm based molecular modeling technique for RNA stem–loop structures

Hiroyuki Ogata, Yutaka Akiyama and Minoru Kanehisa*

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611, Japan

Received October 19, 1994; Revised and Accepted December 20, 1994

ABSTRACT

A new modeling technique for arriving at the three dimensional (3-D) structure of an RNA stem–loop has been developed based on a conformational search by a genetic algorithm and the following refinement by energy minimization. The genetic algorithm simultaneously optimizes a population of conformations in the predefined conformational space and generates 3-D models of RNA. The fitness function to be optimized by the algorithm has been defined to reflect the satisfaction of known conformational constraints. In addition to a term for distance constraints, the fitness function contains a term to constrain each local conformation near to a prepared template conformation. The technique has been applied to the two loops of tRNA, the anticodon loop and the T-loop, and has found good models with small root mean square deviations from the crystal structure. Slightly different models have also been found for the anticodon loop. The analysis of a collection of alternative models obtained has revealed statistical features of local variations at each base position.

INTRODUCTION

RNAs perform a wide variety of biological functions such as self-cleaving reactions of ribozymes, protein synthesis by rRNAs and recognition of aminoacyl-tRNA synthetases by tRNAs. These functions should be understood not only on the basis of their primary sequences and secondary structures, but ultimately on the basis of their three-dimensional (3-D) structures. The analysis of RNA 3-D structures is indispensable to clarify the structure–function relationships and the evolution of RNAs.

X-ray crystallography and NMR spectroscopy are powerful experimental methods for the determination of protein structures at the atomic resolution. Thus far, however, neither method has been powerful enough for the analysis of RNA structures. Only the structures of tRNAs (1–4) and short synthetic RNAs have been determined by X-ray crystallography. Recently the structures of tetraloops (5,6) and a three-nucleotide hairpin loop (7) have been determined by 2-D NMR spectroscopy. Nikonowicz *et al.* reported that they were applying 3-D heteronuclear NMR study to hammerhead catalytic domain (8). Overall, our knowledge on RNA 3-D structures is quite limited.

To cope with this situation, computer modeling is becoming a substitute method for the analysis of 3-D RNA molecular structures [for review see (9)]. Molecular modeling aims to find the 3-D structures of RNA that satisfy the structural information obtained by various methods such as electron microscopy, neutron scattering, low resolution X-ray and NMR analysis, site-directed mutagenesis, crosslinking, chemical and biochemical probing, phylogenetic comparison and secondary structure prediction by free energy minimization. The constructed models are considered to be useful for the design of further experiments.

Manual or interactive modeling techniques have been utilized to propose 3-D models for 16S rRNA (10,11), 5S rRNA (12,13), $tRNA^{Ser}$ (14), group I intron (15–17), U1 snRNA (18) and tetraloops (19). Although the manual approach has been popular, it is dependent on the decisions of experts who construct RNA structural models. Therefore, several automatic modeling techniques have recently been proposed. One employs the distance geometry algorithm to fold pseudo atoms representing RNA molecules (20,21). Another systematically searches the conformational space by building up nucleotides in a discrete nucleotide conformational set (22–24). Conventional molecular mechanics calculations have also been employed for automatic modeling of RNA (25–29). Here we present a new automatic modeling technique suitable for RNA stem–loop structures and its application to the two loops of tRNA.

The main stage of any automatic computer modeling technique is a conformational search procedure or model building phase. We have developed a conformational search program based on a genetic algorithm. Genetic algorithms (GAs), which mimic the natural selection in evolution and efficiently search the combinatorial space (30,31), are recently gaining recognition as an important tool for conformational search of biological macromolecules (32). Lucasius *et al.* (33) used a GA to determine DNA structure from an NMR NOE table. We used their way of variable mapping. Dandekar and Argos (34) reported a GA simulation study on the folding of a four β stranded bundle. GAs have the aspects of both probabilistic and heuristic search algorithms. The efficiency of this duality was demonstrated by Unger and Moult (35) in the study of two dimensional folding simulations by both GA and Monte Carlo simulations. Sun (36) also employed a GA for protein folding prediction with a reduced representation model.

The description of the backbone conformation of a single nucleotide requires six dihedral angles, which is to be compared with only two for an amino acid in the protein backbone. Thus,

* To whom correspondence should be addressed

the conformational space is expected to be enormous, even for a short RNA. Furthermore, it is often the case that the constraints available are too limited to determine a unique structure, so we have to somehow put restrictions to the search space. In fact, manual modeling technique usually limits the conformational space of local segments by choosing one appropriate conformation from a collection of conformations (37). One of the novel aspects of the modeling technique reported in this paper is that the fitness function to be optimized by the GA has a term to stabilize each local segment near to a predefined conformation, in addition to a term to satisfy the distance constraints given by experiments or sequence analysis.

In the present study we focus our analysis on the stem–loop structures which often correspond to functionally important sites. Thus, we adopt an atomic resolution model for local folding patterns of RNA molecules. In comparison, pseudoatom representation might be recommended for global RNA folding problems, such as helix packing of rRNA. Our GA based conformational search technique has been tested for the two loops of tRNA.

MATERIALS AND METHODS

Definition of variables and a segment

In an atomic resolution model of RNA, covalent bond angles and lengths are treated as constant. The conformation of a single nucleotide is defined by the seven variables, α, β, γ, ε, ζ, χ and *P* (Fig. 1). The pseudorotational phase angle *P* describes sugar puckering and defines dihedrals in sugar, v_i, by the following equation (38):

$$v_i = v_{\max} \cos\left[P + \frac{4\pi\ (i-2)}{5}\right], \qquad (1)$$

where $i = 0$–4 and the pucker amplitude is assumed to be constant, $v_{\max} = 38°$. Dihedrals in sugar, v_i, are used for the generation of atomic coordinates. If Cartesian coordinates had been used as variables instead of the above seven variables, there would have been a 9-fold increase in the number of variables required. We analyzed the values of the seven variables that were actually taken in the known RNA structures in the Brookhaven Protein Data Bank (39). As shown in the outside ring of Figure 2 the observed values are localized in some limited ranges. Thus, in the stage of the GA search described below we restrict the sampling of the values to these ranges; specifically, each of the seven variables randomly changes its value in the respective range.

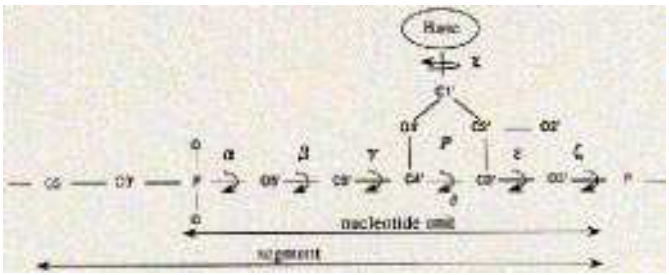

Figure 1. Seven variables describing the nucleotide conformation and the definition of the segment.

It may often be the case in molecular modeling that some parts of RNA are known or assumed to take specific conformations. So, it is necessary to add conformational constraints to local segments of generated models. We define a segment as a part of RNA that is composed of a nucleotide, C3′ and O3′ atoms of the 5′-neighboring nucleotide and P atom of the 3′-neighboring nucleotide (Fig. 1). Two adjacent segments therefore share three atoms.

Local conformations of double helices and single stranded loops

Since double helices of RNAs are found only in the A-form, the conformation of a Watson–Crick base-paired stem is fixed

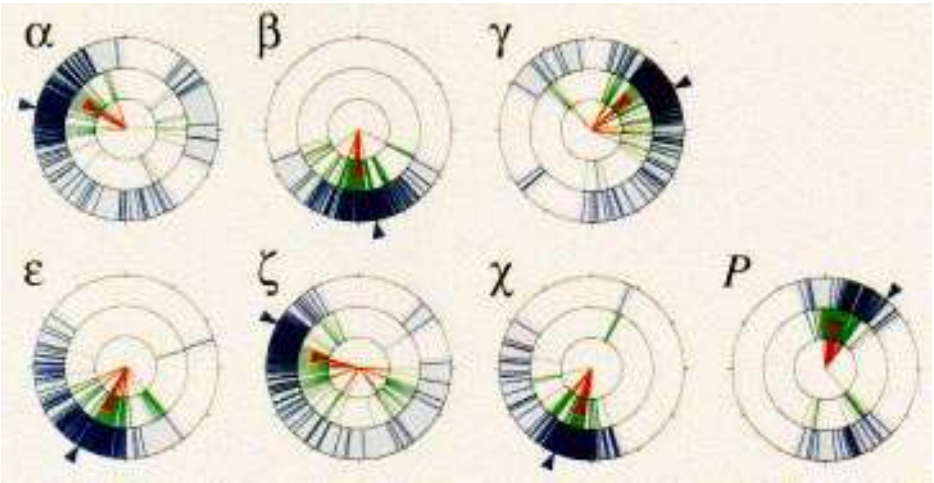

Figure 2. Observed values of the seven variables that are actually taken in the known RNAs (blue outer ring). The data set taken from the Protein Data Bank was composed of 4TNA ($tRNA^{Phe}$), 2TRA ($tRNA^{Asp}$), 1OFX (hybrid RNA/DNA duplex), 1BMV (single stranded RNA from virus), 2TMV (single stranded RNA from virus) and 1RNA (RNA duplex), which included 194 ribonucleotide units. At the stage of the GA search, each variable randomly takes discrete values within the dark and light blue areas. In order to show the actual sampling range, the values taken in the 2000 GA-optimized models and the 72 accepted models for the first segment of the anticodon loop are shown in the green and red inner rings, respectively. The blue and red triangles indicate the values in the crystal structure and the fittest model, respectively.

throughout our modeling process to the idealized A-RNA. The actual A-form conformation was generated with Insight/Discover (Biosym Technologies).

General knowledge about the structure of single stranded loops is not abundant, which is why we focus the search on stem–loop structures. In this study, every segment conformation in a loop is assumed to be stabilized near to A-segment conformation, the conformation of the segment in the ideal A-RNA (see Discussion). It should be noted that even if every segment conformation in a loop is constrained near to A-segment in the search, the conformational space of the entire loop structure is enormous.

Fitness function for the evaluation of model structures

The evaluation function to be optimized is usually called the fitness function in GAs. By gradually optimizing or maximizing the fitness values for a population of conformations, we try to obtain desirable conformations. Our fitness function consists of the three terms:

$$F = F_{dp} + F_{repel} + F_{seg} \quad (2)$$

representing the distance and positional constraints, the stereochemical constraints of atomic collisions and the local conformational constrains over the segments.

The first term of Eq. (2), which evaluates the extent of satisfaction in distance or position constraints, is given by:

$$F_{\mathrm{dp}} = \sum \mathrm{w}_i \mathrm{F}^i_{\mathrm{dp}}\,, \quad (3)$$

and

$$F^i_{\mathrm{dp}} = \begin{cases} 1.0 & (0 \le Diff_i < Diff^u_i) \\ \dfrac{Diff_i - Diff^c_i}{Diff^u_i - Diff^c_i} & (Diff^u_i \le Diff_i < Diff^c_i)\,, \\ 0 & (Diff^c_i < Diff_i) \end{cases} \quad (4)$$

where $Diff_i$ is the violation of the i-th constraint, $Diff_i^{\mathrm{u}}$ and $Diff_i^{\mathrm{c}}$ are the upper limit and the cut-off value of the i-th constraint, respectively, and w_i is the weight for the constraint.

The second term of Eq. (2), a simplified repulsion force, is given by:

$$F_{\mathrm{repel}} = w_{\mathrm{repel}} \sum F^j_{\mathrm{repel}} \quad , \quad (5)$$

and

$$F^j_{\mathrm{repel}} = \begin{cases} \dfrac{r_j}{R_j} & (0 \le r_j < R_j) \\ 1 & (R_j \le r_j) \end{cases} \quad (6)$$

where r_j is the distance between the j-th pair of atoms, R_j is the sum of van der Waals radii of the two atoms and w_{repel} is the weight of this term. The repulsion force is calculated when the two atoms listed in Table 1 are in different nucleotides.

The last term of Eq. (2) enables to restrict each segment conformation. If the k-th segment is assumed to take a known conformation, the atomic coordinates of the k-th segment generated in the process of the conformational search are superposed to the known ones by Kabsch's method (40,41), and the root mean square distance, Rms_k , of all atoms in the segment is calculated. Then F_{seg} is given by:

$$F_{\mathrm{seg}} = \sum w_k \, F^k_{\mathrm{seg}} \quad , \quad (7)$$

and

$$F^k_{\mathrm{seg}} = \begin{cases} \dfrac{Rms^c_k - Rms_k}{Rms^c_k} & 0 \le Rms_k < Rms^c_k) \\ 0 & (Rms^c_k \le Rmx_k) \end{cases} \quad , \quad (8)$$

where $Rms_k{}^{\mathrm{c}}$ is the cut-off value of this term and w_k is the weight for the k-th segment constraint.

Table 1. Atom species for which the repulsion forces are calculated

	Atom species		
main chain and ribose	O5′	O3′	C1′
Adenine	C2	N6	C8
Cytosine	O2	N4	C6
Guanine	N2	O6	C8
Uracil	O2	O4	C6

Genetic algorithm based conformational search

The fitness function defined above is gradually optimized by the search procedure employing the concept of GAs to find desirable conformations. In the field of GAs, candidates of solutions kept in the system is named individuals. Our GA based search algorithm starts with N individuals; that is, the starting population contains N conformations. By randomly changing and mixing the variables describing conformations, a new population of conformations is generated (genetic operation). Then N individuals are selected in accordance with their fitness (selection and creation of new generation). Iterating these two processes, we obtain conformations with higher fitness values.

The conformation of RNA is determined uniquely by the collection of nucleotide conformations, each of which is described by the seven variables mentioned above. In the conformational search, a variable is assumed to take n bits of discrete real values. Thus, when n bits are assigned to each variable, the conformation of RNA consisting of k nucleotides is represented by $7kn$ bit string. This long bit string is called a 'chromosome'. In GAs, a method called Gray coding is often used for a way of mapping between a decimal number and a bit string. If the difference of two decimal numbers is one, the corresponding bit strings can differ by several bits in the binary codes, but they always differ by one bit in the Gray codes. So it is easier for the Gray codes than the binary codes to move around adjacent values.

In each generation of the GA, N chromosomes are kept and they undergo two types of genetic operations: 'mutation' and 'crossover'. A mutation operation corresponds to random changes in variables. N chromosomes are copied as parents and they are subject to mutation operations. First an individual is randomly selected from these N copies. Then a position of its chromosome is randomly selected for a point mutation, which may result in a change of the bit. This process is iterated for m times. After the

mutation operation cycle, the total population size in the pool of chromosomes becomes 2*N*.

A crossover operation is to mix parts of two chromosomes to generate a new one. In this study, one-point crossover is adopted. One pair of chromosomes is selected from 2*N* chromosomes generated by the mutation operation. One crossover point is selected from the joints of the variables assigned on the chromosomes and left or right part of the two parents are exchanged. The fittest of the two sons brought by one crossover operation is added to the pool of chromosomes. The crossover operation is considered to be the most important procedure in GAs, which is absent in other conformation search algorithms such as Monte Carlo simulation and simulated annealing. The crossover operation is iterated *c* times. After this operation, the total population in the pool of chromosomes becomes $2N + c$.

From the pool of chromosomes of parents and sons, *N* chromosomes are selected according to their selection probabilities. The selection probability p_i of the *i*-th chromosome is defined as:

$$p_i = \frac{F_i - F_{\min}}{\sum(F_i - F_{\min})} \quad , \qquad (9)$$

where F_i is the fitness of the *i*-th chromosome and $F_{\min}$ is the fitness of the chromosome that has the least fitness in the pool of chromosomes.

We iterate these genetic mutation, crossover and selection procedures until we obtain conformations with reasonable fitness or until a predescribed number of generations is reached. The computer program is written in C, and it is executed on a Sun workstation.

Refinement by energy minimization

The model conformations obtained by the GA search are then refined by the energy minimization in vacuum to remove steric hindrances. Using CHARMm (42) 200 cycles of steepest descents minimization are followed by 1000 cycles of adopted basis Newton–Raphson minimization with a distance-dependent dielectric constant ($\varepsilon_0 = 4r$). SHIFT function and SWITCH function available in CHARMm are used as smoothing functions of the Lennard–Jones potential and the electrostatic potential, respectively. The stem regions are fixed in all minimization procedures. The calculation is carried out on a supercomputer, CRAY Y-MP2E.

RESULTS

Modeling of tRNA stem–loops

Our technique was tested on the anticodon arm and the T-arm of $tRNA^{Phe}$. Each of the arms consists of a five base-paired stem and a seven nucleotides loop (Fig. 3). For simplicity, modified bases were replaced by their metabolic parents. The anticodon loop is involved in the inter-molecular interaction with mRNA and the T-loop in the intra-molecular interaction with other parts of tRNA. In the case of the anticodon loop modeling, the five nucleotides, G34, A35, A36, G37 and A38 were assumed to be stacked. This structural information had been proposed before the X-ray crystal structure was solved (43). The structural information adopted for the T-loop modeling consisted of the base

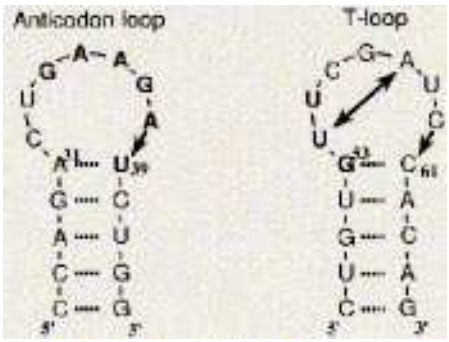

Figure 3. Structural information adopted for the anticodon loop and the T-loop calculations. Stacked bases are indicated by bold letters. ←: positional constraint. ↔: reverse Hoogsteen base-pair. This information is transformed to the constraints between atoms (Table 2) in the conformational search process.

Table 2. Distance constraints representing the structural knowledge

arm	5'-base	3'-base	atom in 5'-base	atom in 3'-base	distance (Å)
anticodon arm	G	A	N1	N7	3.6
			N3	N7	3.5
			C5	N7	3.4
	A	A	N1	N7	3.6
			N3	N7	3.5
(stacking)			C5	N7	3.4
	A	G	N1	N7	3.6
			N3	N7	3.5
			C5	N7	3.4
	A	U	N3	N1	3.6
			N3	N3	4.2
			N3	C5	3.5
T-arm	G	U	N3	N1	3.7
			N3	N3	4.2
			N3	C5	3.4
(stacking)	U	U	O2	N1	3.7
			O2	N3	4.4
			O2	C5	3.5
(reverse Hoogsteen	U	A	O2	N6	2.9
base-pair)			N3	N7	2.9

stacking assumption of the bases G53, U54 and U55, and the reverse Hoogsteen base-pair between U54 and A58. This base-pair could be detected by phylogenetic comparison (44) and low resolution NMR spectroscopy (45,46). All the structural information described here was also used in the loop modeling by Major *et al.* (22).

The knowledge on the base stacking and the reverse Hoogsteen base-pair is incorporated in the first term of Eq. (2) as distance constraints between pairs of atoms (Table 2). Taking into consideration that the stacking pattern is characterized by the polar group of one base superposed over the aromatic system of the adjacent base (38), the atom pairs representing base stacking were selected as shown in Table 2, and the distances were taken from the ideal A-RNA. The constraints concerned with the reverse Hoogsteen base-pair were given from the data of hydrogen-bonding distance determined by neutron diffraction (38). The repulsion force, the second term of Eq. (2), was considered for the atoms in the seven nucleotides of the loop and in the two nucleotides of the stem (A31 and U39 for the anticodon stem; G53 and C61 for the T stem). The position of the P atom of the last nucleotide in the loop is restricted by the position constraints of the first term of Eq. (2).

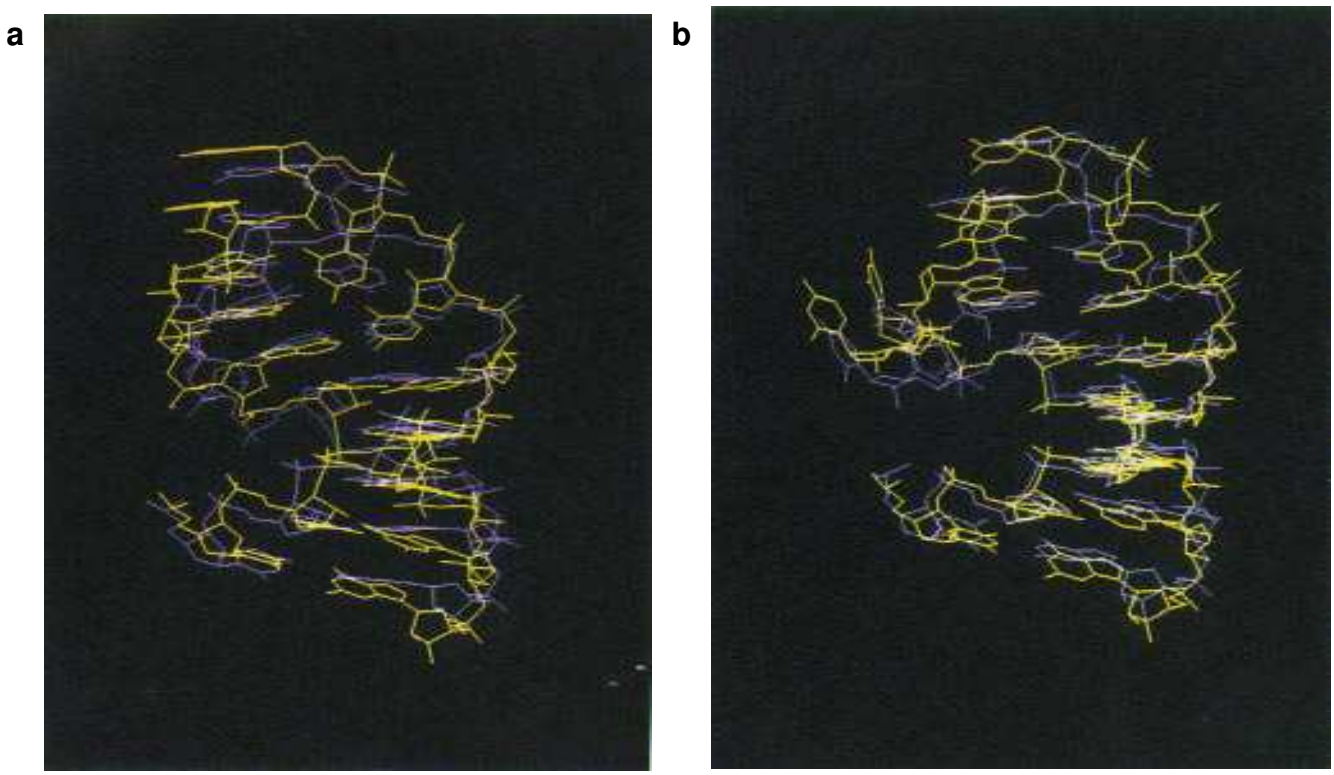

Figure 4. The model structure of the lowest rmsd (yellow) in comparison with the crystal structure (blue): (**a**) the anticodon loop and (**b**) the T-loop. The two pyrimidine bases of U59 and C60, which project out on the left tip of the T-loop, were in agreement with the crystal structure.

The initial structures of the two loops were taken as ideal A-RNAs, which completely satisfied the local conformational constraints as specified by the third term of Eq. (2). Atomic coordinates of the loop were generated successively from the 5′-end of the loop, so for example, A38 and U39 of the anticodon loop were not closed in the initial structure.

Seven bits were assigned to each variable in the chromosome representation. As such, each variable could take 128 different values in the predefined range. Since a nucleotide conformation is defined by seven variables, its conformational space is of the order of 10^{14}. We expect this search space would be sufficient for a rough global search. $Diff_i^u$ and Rms_k^c were set to 0.1 and 2.0 Å, respectively. The sum of the radii of two atoms, R_j, was set to 3.0 Å. With the exception of $Diff_i^u$, Rms_k^c and R_j, the parameter values were adjusted in order to obtain acceptable models (see below). Therefore, $Diff_i^c$ was set to 20 Å. The weights for the three terms of Eq. (2) were 25, 40 and 1, respectively. The population N was set to 20, the number of mutation operations m was 20 (mutation rate was 1 bit/individual × generation) and the number of crossover operations c was 4. One trial of the GA search was set to 2000 generations, and 100 trials with different random number seeds were performed for each of the anticodon loop and the T-loop. The individual with the highest fitness was always selected as a constituent of the next generation (elitist model). The calculation time for one GA search was ~25 c.p.u. min on a SPARCserver 690. 2000 (20 × 100) model structures were obtained after the calculation for each of the two loops.

We then identified acceptable model structures for further refinement analysis by the following three criteria: (i) Only one model was selected from duplicates of the same models. (ii) The root mean square violations of the upper limits $Diff_i^u$ in Eq. (2) and of the sum of van der Waals radii R_j in Eq. (3) must be lower than 0.15 Å. (iii) The distance of up to 2.5 Å was permitted for the loop closure between O3′ of the last nucleotide of the loop and P of the 3′-neighboring nucleotide in the stem.

These criteria reduced the number of model structures to 72 for the anticodon loop and 148 for the T-loop. The accepted models were then subject to refinement by energy minimization.

Comparison with the crystal structure

The refined model that was closest to the crystal structure is shown in Figure 4a for the anticodon loop and in Figure 4b for the T-loop. For the anticodon loop, the root mean square distance (rmsd) was 1.81 Å for all atoms, 1.63 Å for the main chain atoms, 1.52 Å for the phosphorus atoms, and 1.46 Å for the glycosylic nitrogen atoms. This best model ranked at the 22nd in terms of the fitness values of the 72 accepted models; namely it was not the fittest. For the T-loop, the rmsd was 1.76 Å for all atoms, 1.37 Å for the main chain atoms, 1.28 Å for the phosphorus atoms, and 1.42 Å for the glycosylic nitrogen atoms. The best model was the 72nd in terms of the fitness values of the 148 accepted models. For comparison, the algorithm by Major *et al.* found models with the rmsd of 2.00 Å for the anticodon loop and 2.35 Å for the T-loop (22).

Why was the best crystal-like structure not the fittest? First of all, the structural information adopted to this test may not have been adequate to reconstruct the crystal structure. Secondly, the distance constraints in Table 2 may not have been sufficient to describe the structural information. In any event, we consider that the GA based global search must have covered an extensive conformational

Table 3. Comparison of the accepted models and the crystal structure

loop	model	Number of models	Rmsd (Å)			
			All	main chain	P	glycosyl-N
Anticodon loop	Accepted	72	3.81	3.93	3.91	3.10
	Best	1	1.81	1.63	1.52	1.46
	Mean	1	3.24	3.41	3.38	2.55
T-loop	Accepted	348	3.43	3.17	3.19	2.86
	Best	1	1.76	1.37	1.28	1.42
	Mean	1	2.34	2.37	2.40	1.68

space to find a crystal-like structure among the accepted models. The wide sampling range covered by the search can be seen in Figure 2 for the first segment of the anticodon loop in the GA-optimized 2000 model structures and the 72 accepted models.

The mean rmsd between all the energy-minimized models and the crystal structure is presented in Table 3 together with the rmsd for the best model. In addition, we computed the mean structure by averaging the coordinates after superposing all the refined models. The rmsd between this mean structure and the crystal structure is also presented in Table 3. The glycosylic nitrogen atoms had lower rmsd than phosphorus atoms for both the loops. It might be because the distance constraints adopted in the test restrain the atoms in the bases more than those in the main chain. The correlation coefficient between the refined energy value and the rmsd was 0.45 for the anticodon loop and 0.49 for the T-loop.

Table 3 also indicates that the models for the T-loop are closer to the crystal structure than the models for the anticodon loop. The crystal structure of the anticodon loop is actually closer to the A-form than that of the T-loop. The mean rmsd between segments in the crystal structure and the A-form segment is 0.55 Å for the anticodon loop and 0.87 Å for the T-loop. So the A-form constraints over the segments would have been advantageous for the anticodon loop rather than the T-loop. However, the long range constraints of the reverse Hoogsteen base-pair seem to have decreased the flexibility of the conformation of the T-loop. In fact, for the anticodon loop quite different structures that satisfy the five base stacking constraints were found in the models (Fig. 5). One of them had a different stacking pattern, six base stacking on the 3′ strand. A structure with the loop bent over the major groove was also found. In contrast, there were no noticeably different models for the T-loop.

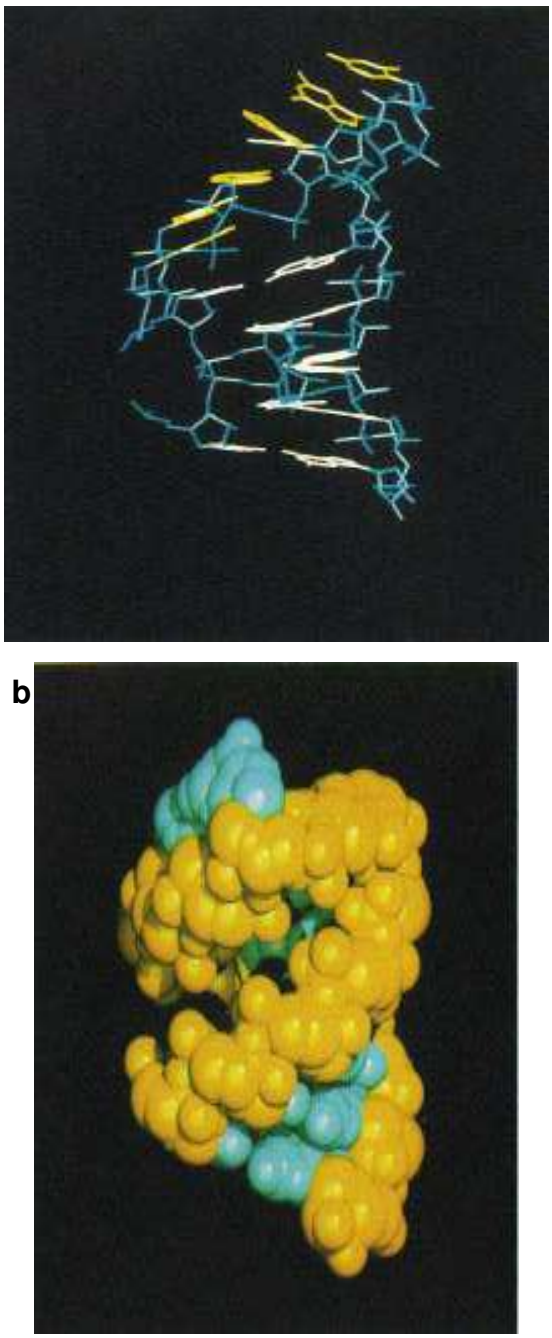

Figure 5. There were some variations of the model structures of the anticodon loop. (**a**) A model for the anticodon loop with six bases stacking from U33 to A38. The model satisfies stacking constraints over the five bases. (**b**) A model for the anticodon arm with the loop bent over the major groove.

Statistics of model structures

The accessible surface area (47) is an index characterizing the environment around atomic groups. In Figure 6a and b the mean and the standard deviation of the accessible surface areas of the bases in model structures are plotted for each base position of the loop compared to the accessible surface area in the crystal structure. Overall, the values were in agreement with the crystal structure. This suggests that the environment of the bases that were not constrained by base stacking or base pairing could be determined from the structural information of other bases. For example, U59 of the T-loop has high accessibility and is exposed in the crystal structure. U59 and C60 take a similar structure of bulges that loop out of the duplex, because of the adjacent base-pairs of C61–G53 (Watson–Crick) and A58–U54 (reverse Hoogsteen). No distance constraints were adopted for this pyrimidine. The predicted mean accessible surface area is close to that of crystal structure, although the value for C60 is overestimated. The prediction for C56 and G57, which have no distance constraint, are also good.

In order to estimate the fluctuation or ambiguity of the atomic coordinates, the root mean square deviations from the mean coordinates were computed for each of the P atoms and the glycosylic nitrogen atoms (Fig. 6c and d). The values were relatively small for the bases with stacking and/or base-pairing constraints: A35, A36, G37 and A38 in the anticodon loop and U54, U55 and A58 in the T-loop. The base G34 of the anticodon

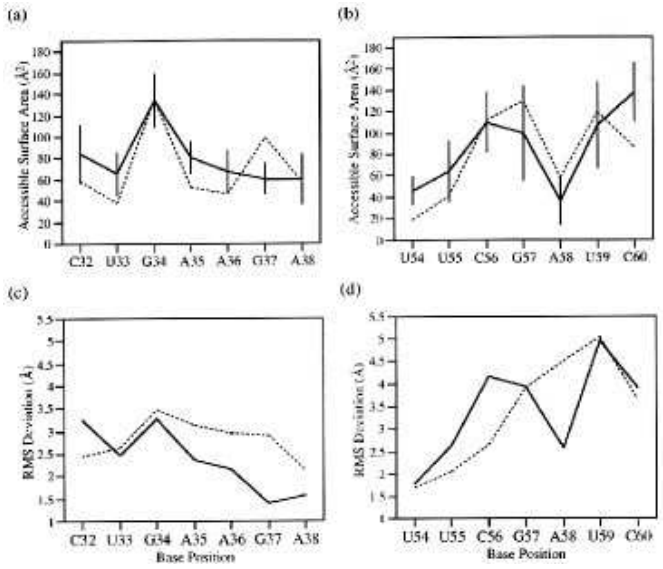

Figure 6. The accessible surface area of the bases and the root mean square deviations of glycosyl nitrogen and phosphorus. In the upper panels, the accessible surface area is plotted versus each base position for (**a**) the anticodon loop and (**b**) the T-loop. The mean and standard deviation of the model structures (solid line) in comparison with the crystal structure (dotted line). In the lower panels, the root mean square deviations of the coordinates of glycosyl nitrogen (solid line) and phosphorus (dotted line) from the mean coordinate are plotted for (**c**) the anticodon loop and (**d**) the T-loop. Five base stacking (G34, A35, A36, G37, A38) in the anticodon loop, and two base stacking (U54, U55) and a base-pair between U55 and A58 in the T-loop were assumed in the calculations.

loop, which had stacking constraints, was more ambiguous than the other bases with stacking constraints. This is consistent with the finding of a bent loop model (Fig. 5b). This indicates that the local constraints of the base stacking alone could not determine the global conformation of the anticodon loop and, conversely, the information of long range interaction was important for the determination of local structure.

DISCUSSION

A single stranded region of RNA such as a hairpin loop or a bulge loop often participates in the interaction with other molecules. The loop structure tends to be strongly dependent on its sequence, in comparison with the Watson–Crick base-paired region which tends to be standard helical A-RNA. Different structural elements formed by loops may be responsible for functional diversity and specificity of RNAs. For example, coat proteins of R17 and Qβ have specific interactions with RNA hairpins (48,49). The HIV *rev* protein recognizes RNA stem–loop and activates specific gene expression (50). Also the HIV *tat* protein interacts with *tar* RNA, which contains hairpin and bulge, and activates the expression of HIV genes (51). A 3′-terminal stem–loop of histone mRNA is considered to be essential for the post-transcriptional coupling of histone mRNA levels to DNA synthesis in mouse fibroblasts (52). Many conserved tetraloop caps are observed in rRNA (53). The analysis of loop structures may throw light on the general properties of RNA structures and the expression of RNA functions.

We proposed a technique for the 3-D modeling of RNA stem–loops. The conformational space that satisfy the given structural information is searched by a GA. Because the search space can be very extensive even for short RNAs and the target function to be optimized is usually multimodal, the multiple minima problem is inevitable. In the modeling of two loops of tRNA, each of which consists of seven nucleotides, 7 bits were assigned to each of the seven variables. Since each variable can take 2^7 different values, the search space is about 10^{100}. GAs have been considered more promising to overcome the multiple minima problem than other probabilistic search algorithms, which we consider to be confirmed by the successful application to the 3-D molecular modeling of tRNA loops.

In this study, a conformation of an RNA was defined by the internal coordinates: six dihedral angles and the pseudorotational phase angle. The Cartesian coordinates of atoms can define a conformation of an RNA as well. However, the choice of the internal coordinates is better for the GA because random changes of variables by the genetic operation often result in unrealistic bond lengths and bond angles when the position vectors are the variables describing a conformation. Moreover, the choice of internal coordinates system reduces the number of variables by one ninth.

The point mutation operation is implemented here to change the variables in the internal coordinate system, which then affects the conformation of an RNA. Instead, the conformation of a segment could be the direct target of mutation operations. This approach has an advantage of reducing the search space, when variables describing the conformation are highly correlated and also if different conformations can be properly weighted. However, we did not take this approach. Hundreds of X-ray crystal structures have been determined for proteins, and the strong correlation is observed between ϕ and ψ in the Ramachandran map. In contrast, the structural knowledge about nucleotide units in large RNAs is very limited, since only a handful of RNA structures have been determined and since the nucleotide backbone unit has the freedom comparable to three amino acid residues. Thus, it is more difficult to define a set of favorable conformations for nucleotide units than for amino acid residues. Exploring the preference of nucleotide conformations remains important and interesting work.

The basic idea of constructing the fitness function is automating experts' knowledge and manipulation. Our fitness function contains two types of constraints: the global constraint on inter-atomic distances and the local conformational constraint over segments by template superposition. Although the latter constraint may be transformed to the former, it is more convenient to have it separately. In manual modeling, an expert will build up RNA models by consecutively choosing an appropriate segment from a collection of conformations and adding small changes to the conformation selected. These changes are necessary for larger sampling of the search space; if segments attached are rigid, sampling will be very sparse. The third term of Eq. (2) mimics this manipulation not by hand but by a numerical form. The transformation of experts' knowledge on structures into atomic distance representation is intricate and laborious. We represented the structural information by the first term of Eq. (2) and Table 2. This should be considered a first step toward identifying structural data and representing them in numerical forms. Further development remains to be our future work.

The segments in the anticodon loop and the T-loop were assumed to take A-segment-like structures and A-segment constraints were applied in this study, because a lot of works support this assumption. From the analysis of crystal structures of nucleotides, the similarity with the local structure of A-RNA was indicated (54). Single stranded RNAs have long been known to take A-RNA-like

conformation in aqueous solution (55). Since apurinic acids also have the same property, the main chain interaction is considered to be responsible for the right handed helical propensity of nucleic acids rather than the base stacking interaction (56). With this background, Kajava pointed out that A-RNA was appropriate for the initial structure for modeling of tetraloops (19). When a single stranded region of RNAs has interaction with other molecules *in vivo*, its local structure may not take A-RNA-like structure. The analysis of crystal structures of nucleotide–protein complexes, however, revealed that the conformation of the nucleotides is not significantly different from the free state (57).

The GA based search is rather rough and does not aim at a local search. In fact, the GA accepted models sometimes had steric hindrances because of the simplicity of the fitness function and the nature of non-differential search of the algorithm. Therefore, we combined the global search by the GA with a local search by the energy minimization. This strategy of combining global and local searches is not new, but we think our results were obtained largely by the new global search algorithm. The results indicate that the structural information adopted for the T-loop modeling was sufficient to predict the atomic coordinates, at least, of the glycosylic nitrogen atoms. In addition, the analysis of alternative models revealed the local variation of each base as shown in Figure 6. It is one of the purposes of the automatic molecular modeling to get this kind of statistical information. It is difficult to obtain statistical features from manual modeling which handles one single structural model at a time.

The technique reported here is able to deal with other types of structures such as bulge loops and pseudoknots. Because of the limitations of computational resources currently available, we think the GA based algorithm with an extensive sampling space as reported in this paper is applicable only to loops shorter than about 20 nucleotides. However, if long-range structural information is available for a larger molecule, the global search problem can be divided into small problems as done by the entire tRNA modeling by Major *et al.* (24). Further improvements of the technique will also lead us to the modeling of larger RNAs that play various roles in a cell.

ACKNOWLEDGEMENTS

This work was supported in part by the grant-in-aid for scientific research on the priority area 'Genome Informatics' from the Ministry of Education, Science and Culture. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

REFERENCES

1 Kim,S.-H., Quigley,G.J., Suddath,F.L., McPherson,A., Sneden,D., Kim,J.J., Weinzierl,J. and Rich,A. (1973) *Science*, **179**, 285–288.
2 Moras,D., Comarmond,M.B., Fischer,J., Weiss,R., Thierry,J.C., Ebel,J.P. and Giege, R. (1980) *Nature*, **288**, 699–674.
3 Woo,N.H., Roe,B.A. and Rich,A. (1980) *Nature*, **286**, 346–351.
4 Rould,M.A., Perona,J.J., Soll,D. and Steitz,T.A. (1989) *Science*, **246**, 1135–1142.
5 Cheong,C., Varani,G. and Tinoco,I.Jr. (1990) *Nature*, **346**, 680–682.
6 Heus,H.A. and Pardi,A. (1991) *Science*, **253**, 191–194.
7 Davis,P., Thurmes,W. and Tinoco,I.Jr. (1993) *Nucleic Acids Res.*, 21, 537–545.
8 Nikonowicz,E.P. and Pardi,A. (1992) *Nature*, **355**, 184–186.
9 Gautheret,D. and Cedergren,R. (1993) *FASEB J.*, **7**, 97–105.
10 Brimacombe,R., Atmadja,J., Stiege,W. and Schuler,D. (1988) *J. Mol. Biol.*, **199**, 115–136.
11 Stern,S., Weiser,B. and Noller,H.F. (1988) *J. Mol. Biol.*, **204**, 447–481.
12 Westhof,E., Romby,P., Romaniuk,P.J., Ebel,J.-P., Ehresmann,C. and Ehresmann,B. (1989) *J. Mol. Biol.*, **207**, 417–431.
13 Brunel,C., Romby,P., Westhof,E., Ehresmann,C. and Ehresmann,B. (1991) *J. Mol. Biol.*, **221**, 293–308.
14 Dock-Bregeon,A.C., Westhof,E., Giege,R. and Moras,D. (1989) *J. Mol. Biol.*, **206**, 707–722.
15 Michel,F. and Westhof,E. (1990) *J. Mol. Biol.*, **216**, 585–610.
16 Jaeger,L., Westhof,E. and Michel,F. (1991) *J. Mol. Biol.*, **221**, 1153–1164.
17 Benedetti,G. and Morosetti,S. (1991) *J. Biomol. Struct. Dyn.*, **8**, 1045–1055.
18 Krol,A., Westhof,E., Bach,M., Luhrmann,R., Ebel,J.-P. and Carbon,P. (1990) *Nucleic Acids Res.*, **18**, 3803–3811.
19 Kajava,A. and Ruterjans,H. (1993) *Nucleic Acids Res.*, **21**, 4556–4562.
20 Hubbard,J.M. and Hearst,J.E. (1991) *Biochemistry*, **30**, 5458–5465.
21 Hubbard,J.M. and Hearst,J.E. (1991) *J. Mol. Biol.*, **221**, 889–907.
22 Major,F., Turcotte,M., Gautheret,D., Lapalme,G., Fillion,E. and Cedergren,R. (1991) *Science*, **253**, 1255–1260.
23 Gautheret,D., Major,F. and Cedergren,R. (1993) *J. Mol. Biol.*, **229**, 1049–1064.
24 Major,F., Gautheret,D. and Cedergren,R. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 9408–9412.
25 Mei,H.-Y., Kaaret,T.W. and Bruice,T.C. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 9727–9731.
26 Malhotra,A., Tan,R.K -Z. and Harvey,S.C. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 1950–1954.
27 Veal,J.M. and Wilson,W.D. (1991) *J. Biomol. Struct. Dyn.*, **8**, 1119–1145.
28 Yao,S. and Wilson,W.D. (1992) *J. Biomol. Struct. Dyn.*, **10**, 367–387.
29 Gabb,H.A., Harris,M.E., Pandey,N.B., Marzluff,W.F. and Harvey,S.C. (1992) *J. Biomol. Struct. Dyn.*, **9**, 1119–1130.
30 Holland,J.H. (1975) Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor.
31 Goldberg,D.E. (1989) Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, New York.
32 Wodak,S.J. and Rooman,M.J. (1993) *Current Opinion in Structural Biology*, **3**, 247–259.
33 Lucasius,C.B., Blommers,M.J.J., Buydens,L.M.C. and Kateman,G. (1991) In Davis,L. (ed.), Handbook of Genetic Algorithms. Van Nostrand Reinhold, pp. 251–281.
34 Dandekar,T. and Argos,P. (1992) *Protein Eng.*, **5**, 637–645.
35 Unger,R. and Moult,J. (1992) *J. Mol. Biol.*, **231**, 75–81.
36 Sun,S. (1993) *Protein Science*, **2**, 762–785.
37 Westhof,E., Romby,P., Ehresmann,C. and Ehresmann,B. (1990) In Beveridge,D.L. and Lavery,R. (ed.), Theoretical Biochemistry & Molecular Biophysics. Adenine Press, pp. 399–409.
38 Saenger,W. (1984) Principle of Nucleic Acid Structure, Springer-Verlag, New York.
39 Bernstein,F.C., Koetzle,T.F., Williams,G.J.D., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–545.
40 Kabsch,W. (1976) *Acta Cryst.*, **A32**, 922–923.
41 Kabsch,W. (1978) *Acta Cryst.*, **A34**, 827–828.
42 Nilsson,L. and Karplus, M. (1986) *J. Comput. Chem.*, **7**, 591–616.
43 Fuller,W. and Hodgson,A. (1967) *Science*, **215**, 817–821.
44 Klingler,T.M. and Brutlag,D.L. (1993) In *Proceedings of First International Conference on Intelligent Systems in Molecular Biology*. pp. 225–233.
45 Amano,M. and Kawakami M. (1992) *Eur. J. Biochem.*, **210**, 671–681.
46 Chu,W.-C., Kintanar,A. and Horowitz,J. (1992) *J. Mol. Biol.*, **227**, 1173–1181.
47 Richmond,T.J. (1984) *J. Mol. Biol.*, **178**, 63–89.
48 Romaniuk,P.J., Lowary,P., Wu,H.-N., Stormo,G. and Uhlenbeck,O.C. (1987) *Biochemistry*, **26**, 1563–1568.
49 Witherell,G.W. and Uhlenbeck,O.C. (1989) *Biochemistry*, **28**, 71–76.
50 Olsen,H.S., Nelbock,P., Cochrane,A.W. and Rosen,C.A. (1990) *Science*, **247**, 845–848.
51 Feng,S. and Holland,E.C. (1988) *Nature*, **334**, 165–167.
52 Levine,B.J., Chodchoy,N., Marzluff,W.F. and Skoultchi,A.I. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 6189–6193.
53 Woese,C.R., Winker,S. and Gutell,R.R. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 8467–8471.
54 Sundaralingam,M. (1969) *Biopolimers*, **7**, 821–860.
55 Gulik,A., Inoue,H. and Luzzati,V. (1970) *J. Mol. Biol.*, **53**, 221–238.
56 Achter,E.K. and Felsenfeld,G. (1971) *Biopolymers*, **10**,1625–1634.
57 Moodie,S.L. and Thornton,J.M. (1993) *Nucleic Acids Res.*, **21**, 1369–1380.