A Computer Modeling Method for the Three-dimensional Structure of RNA

Hiroyuki Ogata[1] Yutaka Akiyama[1] Minoru Kanehisa[1]
ogata@kuicr.kyoto-u.ac.jp akiyama@kuicr.kyoto-u.ac.jp kanehisa@kuicr.kyoto-u.ac.jp
Institute for Chemical Research
Kyoto University
Uji, Kyoto 611

Abstract

We are developing a computational method for automatically organizing collections of structural knowledge of RNA into a three-dimensional (3-D) form. The goal of our method for modeling of RNA structure is to find, as much as possible, conformations of RNA which satisfy the constraints from experiments and sequence analysis and, at the same time, whose local conformations are close to some representative conformations. For efficient conformational search, we used a genetic algorithm as a trial. We applyed our method in modeling a single stranded region of an RNA for the estimation of efficiency of our method.

Introduction

The biological functions of RNA such as the enzymatic mechanism of ribozymes, the role of rRNA in protein synthesis, and the recognition of tRNA by aminoacyltransferase should be understood not only on the basis of their secondary structures, but ultimately on the basis of their three-dimensional (3-D) structures. Thus far, X-ray crystallography and NMR, which are very powerful experimental methods for the determination of protein structures, have not been so successful in the case of RNA. To cope with this situation, 3-D models have been proposed for Ser-tRNA [1], 16S rRNA [2], 5S rRNA [3,4], U1 snRNA [5] and *Tetrahymena* group I intron [6] using the information of secondary and tertiary interactions given by experiments such as site-directed mutagenesis and closslinking, and by sequence analysis such as phylogenetical comparison and secondary structure prediction. Such structural models of RNA have been critical in clarifying structure-function relationships and in designing subsequent experiments. But most of these models are proposed by the method called interactive modeling [7,8], which is dependant on decisions of experts who construct RNA structural models. In contrast, two procedures have been proposed

[1]京都大学化学研究所, 〒611 京都府宇治市五ヶ庄

as automatic modeling methods. One uses the distance geometry algorithm and folds pseudo atoms [9,10]. Another one systematically searches the conformational space by building up nucleotides in a discrete nucleotide conformational set [11,12]. We are currently developing a third method.

The description of the backbone conformation of a single nucleotide requires six dihedral angles, which is to be compared with only two dihedral angles for the backbone conformation of an amino acid in a protein. Thus, the conformational space is expected to be enormous, even for a short RNA. Since we have only limited information available for determination of a stable structure, we have to somehow put restrictions to the search space. Model structures are helpful when there are possible structures which are consistent with the given structural information. For the search to be feasible, local conformations of RNA are restricted near to some representatives, which are taken as energetically favorable conformations and/or frequently observed conformations in known RNAs. It is often the case that structural information about RNA is represented as distance constraints. Thus, the problem here is how to find out model conformations which satisfy not only the distance constraints given by experiments or sequence analysis but also the conformational restrictions on each local segment. In the present study, we define a local segment as a part of RNA which is composed of a nucleotide, C3' and O3' atoms of the 5'-neighboring nucleotide and P atom of the 3'-neighboring nucleotide. Two adjacent segments thus share three atoms, C3', O3' and P. We restrict, but not strictly, the conformational search space of segments near some representatives. We then apply genetic algorithms for an efficient conformational search, which are recently gaining recognition as an important conformational search technique [13,14]. We test our method by modeling a hypothetical hepta-nucleotide loop of an RNA.

Method

1) variables

In our method, bond angles and bond lengths are considered to be constant. We use the following equation to define torsions in sugar, $\nu 0$, $\nu 1$, $\nu 2$, $\nu 3$ and $\nu 4$,

$$\nu_i = \nu_{max} \cos\left[P + \frac{4\pi\ (i-2)}{5}\right],$$

where P is psuedorotational phase angle and ν_{max} is pucker amplitude [15]. Pucker amplitude ν_{max} is assumed to be constant. Then each nucleotide has six torsions, α, β, γ, ε, ζ and χ, and one psuedorotational phase angle, P. Thus, a conformation of one nucleotide is defined by seven variables. We restrict the range of possible values for

some variables: β moves from 90° to 270°, ε from 150° to 300° and P from -60° to 40° and also from 140° to 210°. Other four variables take all values between 0° to 360°.

2) genetic algorithm

We employ a genetic algorithm which, in a sense, mimics the natural selection in evolution and efficiently searches the combinatorial space [16]. Each variable or angle is represented by a gray coded bit string in our genetic algorithm. In genetic algorithms, gray coding is often used for a way of mapping between decimal number and bits. A conformation of RNA is determined by the set of the variables mentioned above and is represented by a long bit string called a chromosome. We use n bits for a variable, so that each variable can take 2^n possible values. $7n$ bits are necessary for a nucleotide, because one nucleotide has seven variables. If, for example, eight bits are assigned to one variable, a chromosome representing an RNA of tetra-nucleotide requires 8×7×4=224 bits long. Fitness of each chromosome or RNA conformation should be defined to reflect the degree of satisfaction to available information about conformation of RNA (see below). We can obtain structures which best satisfy constraints via optimizing the fitness of chromosomes.

Our genetic algorithm method starts with N individuals. That is, the starting population of bit strings, or chromosomes, is N. In each generation of the genetic algorithm, N chromosomes are kept and they undergo two types of operations. First, certain chromosomes are selected as parents for next generation and subject to point mutations which correspond to random changes in variables of RNA conformation. Second, certain pairs of chromosomes are selected for crossover operations. One crossover operation brings two sons. The crossover operation is considered to be the most important procedure in genetic algorithms, which is absent in other search algorithms such as Monte Carlo simulation and simulated annealing. From the pool of chromosomes of parents and sons, N chromosomes are selected according to again their selection probabilities. The selection probability p_i of i-th chromosome is defined as

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

where F_i is the fitness of i-th chromosome and F_{min} is the fitness of a chromosome which has the least fitness in the generation. We iterate this procedure until we obtain chromosomes with a reasonable fitness, which may be our desirable conformations.

3) fitness function

The fitness of each chromosome or RNA conformation is defined as

$$F = \sum F_l + \sum F_g.$$

Fl corresponds to the extent of satisfaction in conformational restriction of each local segment. Fg corresponds to the extent of satisfaction in each global constraint. In our current method, we have to decide, before starting a conformational search, to which representative conformation each local segment should be close. A chromosome is temporally transformed to Cartesian coordinates representing a 3-D conformation, and each local segment is best fitted to a corresponding predetermined representative. We use r.m.s.distance d after best fitting, as a similarity measure between a representative and a generated local conformation. Thus we simply define Fl as Fig. (1).

We treat such a type of global structural information that can be represented as constraints on the distance between two atoms. The difference c between the distance in the generated conformation and desirable one is calculated. Then Fg is defined like Fl as Fig.(2).

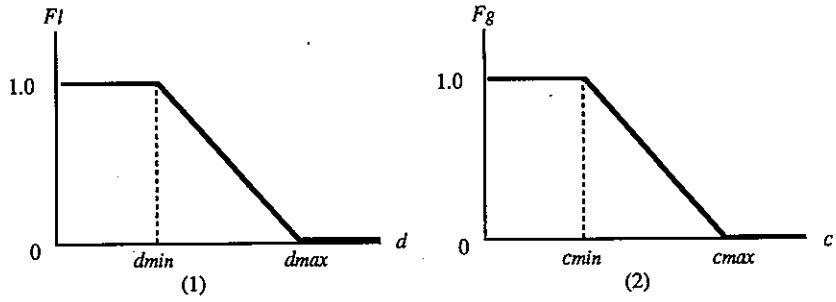

Figure: Definition of Fl and Fg

Results and Discussion

We tested our method by constructing a single stranded hepta-nucleotide (C-U-G-A-G-A-A) loop. In this trial we simply assumed that each segment tended to have two conformations, a and b. One is the constituent part of A-form RNA: a, and the other is B-form: b. We also supposed that our only knowledge about global constraint was the loop closure; we only know the coordinate of P atom of the first nucleotide of the hepta-nucleotide and the coordinate of O3' atom of the last nucleotide. We fixed the first atom and calculated the distance c between the generated coordinate and the known coordinate of the last atom. Since we have two local representative conformations and build a hepta-nucleotide, the total number of combinations for nucleotide conformations is 2^7=128. Before starting our conformational search, we need to select a combination of representatives from these 128 possibilities. In fact, we executed our algorithm for 128 times with all combinations of two representatives in the hepta-nucleotide. We set $dmin$ = 0.3Å, $dmax$ = 2.0Å, $cmin$ = 0.2Å and $cmax$ = 5.0Å.

The population N was set to thirty and the number of bits assigned for a variable, n, was seven. We set the maximum generation 500 for every 128 search.

With the parameter values given above, our genetic algorithm method found that five combinations of representatives, out of 128 trial searches, reached the optimal fitness value of 8.0. The total cpu time required was about 6 hours on a workstation (SPARC station 690). The five combinations with the optimal fitness are: *(baaabab), (bbabbab), (bbbbabb), (aabbbbb)* and *(bbbbbbb)*. It is possible that other combinations may have optimal fitness satisfying all restrictions and constraints, if we continue conformational search. If we had not been able to find out any conformation which satisfied all restrictions and constraints, it would not have meant that such a conformation does not exit. But our method did find out five conformations in reasonable time. This test problem may have been too simple, but at least we could confirm that the algorithm works. In actual cases there will be a number of difficulties to overcome, such as defining combinations of representatives and making the algorithm more efficient. The search efficiency will probably become better, if we can find optimal settings of many parameters in genetic algorithms.

We are in the process of developing computational tools for optimizing parameters in our method, and testing different types of constraints. Once we can improve the efficiency of our method, we will then approach the problem of using multiple restrictions for local conformations and also determine representatives themselves.

This work is 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 the Chemical Research, Kyoto University.

References

[1] Dock-Bregeon, A. C., Westhof, E., Giege, R. and Moras, D.: J.Mol.Biol., 206, 707-792 (1989)
[2] Stern, S., Weiser, B. and Noller, H. F.: J.Mol.Biol, 204, 447-481 (1988)
[3] Westhof, E., Romby, P., Romaniuk, P.J., Ebel, J.-P., Ehresmann, C. and Ehresmann, B.: J.Mol.Biol., 207, 417-431 (1989)
[4] Brunel, C., Romby, P., Westhof, E., Ehresmann, C. and Ehresmann, B.: J.Mol.Biol, 221, 293-308 (1991)
[5] Krol, A., Westhof, E., Bach, M., Luhrmann, R., Ebel, J.-P. and Carbon, P.: Nucleic Acids Research, 18, 3803-3811 (1990)
[6] Michel, F. and Westhof, E.: J.Mol.Biol, 216, 585-610 (1990)
[7] Westhof, E., Romby, P., Ehresmann, C. and Ehresmann, B.: Theoretical Biochemistry & Molecular Biophysics, 399-409 (Adenine Press, 1990)
[8] Gautheret, D., Major, F. and Cedergren, R.: Method in Enzymology, 183, 318-330 (1990)
[9] Hubbard, J. M. and Hearst, J. E.: Biochemistry, 30, 5458-5465 (1991)
[10] Hubbard, J. M. and Hearst, J. E.: J.Mol.Biol., 221, 889-907 (1991)
[11] Major, F., Turcotte, M., Gautheret, D., Lapalme, G., Fillion, E. and Cedergren, R.: SCIENCE, 253, 1255-1260 (1991)
[12] Gautheret, D., Major, F. and Cedergren, R.: J.Mol.Biol. 229:1049-1064 (1993)
[13] Danderkar, T. and Argos, P.: Protein Engineering, 5, 637-645 (1992)
[14] Unger, R. and Moult, J.: J.Mol.Biol., 231, 75-81 (1993)
[15] Saenger, W.: Principles of Nucleic Acid Structure (Springer-Verlag, New York, 1984)
[16] Forrest, S.: SCIENCE, 261, 872-878 (1993)