A fast genetic algorithm for RNA secondary structure analysis

I. I. Titov, D. G. Vorobiev, V. A. Ivanisenko, and N. A. Kolchanov*

Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, 10 prosp. Lavrent´eva, 630090 Novosibirsk, Russian Federation. Fax: +7 (383 2) 33 1278. E-mail: titov@bionet.nsc.ru

A fast genetic algorithm GArna for mass calculations of RNA secondary structures through the Internet is proposed. The algorithm GArna was used to study the effects of nucleotide composition on characteristics of the secondary structure of random RNA sequences. A contextual characteristics for evaluation of the stability was proposed and the application of standard statistical tests for heterogeneous RNA samplings was justified. The structure-contextual characteristics by which the 5´-untranslated regions of high- and low-expression genes of dicot plants and mammals differ were found, and the results were interpreted in terms of secondary structure influence on translation initiation and on the general scheme of expression regulation. The application of the results obtained for the development of computer methods for RNA structural genomics, in particular, for RNA search in genome sequences, is discussed.

Key words: RNA, secondary structure, translation, untranslated regions, structural genomics, genetic algorithm.

Ab initio calculation of RNA secondary structure is among the primary challenges faced by molecular biology and, in particular, by bioinformatics. The most popular approach is minimization of the structure energy.[1,2] The mfold algorithm[3] makes it possible to calculate several lowest-energy structures through the Internet or using a personal computer. The partition function algorithm[4] calculates the probabilities of formation of complementary pairs averaged over the equilibrium ensemble of secondary structures. The kinetic approach is based on simulation of the formation of RNA secondary structure.[5,6] The secondary structure of RNA obtained in this way does not necessarily match the global energy minimum.

In recent years, RNA structure has been analyzed using the genetic algorithm (GA), suitable for calculation of the conformation of the hairpin at the atomic level of resolution[7] and the secondary structures of longer molecules.[8—12] Historically, the GA has been proposed as a potent method for minimization making use of the analogy with natural evolution (see a review[13]). One advantage of GA is that it is able to search for intermediate states of the RNA folding; therefore, later, it has been classified as a kinetic method. The genetic algorithm is based on the assumption that a combination of two "good" solutions of a problem may produce a better solution (so-called building block hypothesis[14]). Therefore, due to the additivity of parametrization of secondary structure energies,[15] application of the GA to the search for RNA folding is quite natural. One more important advantage of the GA is absolute freedom in choosing the functional form of this parameterization, so that tertiary interactions, for example, pseudoknots, can also be taken into account.[8] However, the GAs for prediction of the RNA secondary structure known to date require dozens of hours of personal computer work[8] or the use of supercomputers with parallel architecture[10] and, therefore, they cannot yet compare with dynamic algorithms in computation speed. Therefore, currently, there is no GA that could solve the problem in question through the Internet.

Here we present the first genetic algorithm GArna, whose speed of operation allows one to identify the optimal (or nearly optimal) secondary structure of RNA through the Internet or to perform mass calculations on a PC. An increase in the productivity is attained due to representativeness of the initial population, recursive computation of the secondary structure energy, symmetrization of recombinations, and local optimization. Using GArna, we studied the dependence of the energy distribution of the secondary structure on the nucleotide composition, proposed a contextual characterization of the secondary structure stability, and carried out comparative analysis of 5´-untranslated regions (UTR) in the mRNA of genes of dicot plants and mammals. The approaches proposed in this work can provide the basis for the development of procedures for the computer search for RNA.

Published in Russian in *Izvestiya Akademii Nauk. Seriya Khimicheskaya,* No. 7, pp. 1047—1056, July, 2002.

1066-5285/02/5107-1135 $27.00

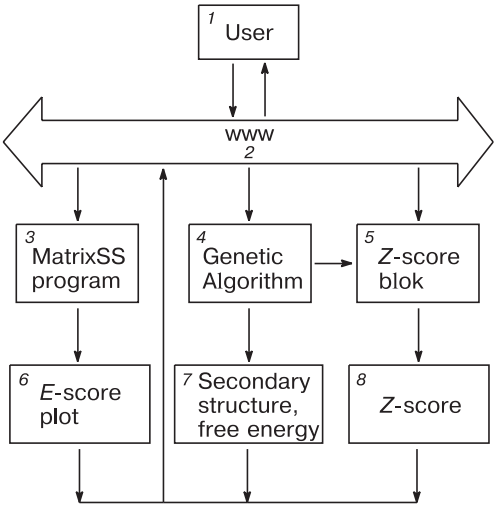

Fig. 1. Diagram of the system for RNA secondary structure analysis based on the GArna genetic algorithm: (*1*) user; (*2*) Web interface; (*3*) MatrixSS program (calculates the complementarity index *E*-score for 250 to 10000 nt long sequences); (*4*) the program running the genetic algorithm (calculates the low-energy secondary structures of RNA sequences with lengths of up to 250 nt); (*5*) block for the calculation of the relative stability *Z*-score for RNA with lengths from 50 to 250 nt; (*6*) plot output unit; (*7*) secondary structure, free energy; (*8*) output of *Z*-score.

System and Methods

The GArna system is available at the URL* and contains the following functional blocks (Fig. 1).

1. The program realizing the genetic algorithm, which calculates low-energy secondary structures of RNA sequences with a length of up to 250 nucleotides (nt).

2. The unit for calculation of the relative stability of RNA with a length of 50 to 250 nt (*Z*-score).

3. The MatrixSS program, which calculates the complementarity index *E*-score for 250 to 10000 nt long sequences.

4. The Web interface.

Programs. The programs were written in the C ANSI system, compiled on an Intel PC platform, and installed to run under Windows NT 4.0.

Energy rules. When calculating the secondary structure energy, parametrization was used.[15]

Sequences. Several types of sequences were used in the study**.

A. The tRNA and 5S RNA sequences were extracted from the EMBL database.[17]

B. The 5′-UTR sequences of mRNA of dicots and mammals were taken from the EMBL database.[17] The sampling was formed according to the principle of division into high- (H) and low-expression (L) genes proposed in the literature.[16] The TRANSFAC database was employed to select the mRNA that encode the transcription factors.[17]

C. The human mRNA 5′-UTR sequences have been previously used for computer analysis,[18] and they were kindly provided by the authors. From this sampling, we eliminated sequences with lengths of less than 50 nt, while in long sequences, the first 250 nt were taken for the analysis. The resulting sequences were split into H and L groups, according to the principle mentioned above.[16]

Algorithm

The genetic algorithm simulates the evolution of a population of artificial "individuals". Each individual has its own set of "genes". The genes determine the organism characters on whose basis the fitness of an individual is evaluated and selection is done. The population evolves as a result of cyclic action of so-called genetic operators: (i) selection through fitness assessment, (ii) recombinations that perform large-scale search, and (iii) mutations that prevent the premature convergence.

In the case of optimization of RNA secondary structure, an individual corresponds to a definite structure variant. The secondary structure of RNA is determined unambiguously by a set of double-stranded sections (helices) forming it; therefore, the GA takes the helices as genes.[8,12] This choice has a natural motivation based on the assumption that the GA simulates the process of RNA folding into a secondary structure by consecutive formation and destruction of helices.[19] With such consideration, the genetic operators of our algorithm bear the following analogies with natural processes (Fig. 2):

— mutations simulate local changes in the RNA secondary structure, *i.e.*, those caused by the formation/disruption of a limited number of helices;

— recombinations perform the exchange by large fragments of the RNA secondary structures;

— selection rejects less stable secondary structures (these structures normally consist of a smaller number of base pairs[20]) and, therefore, it simulates structure degradation.

The operation of GArna consists of the following procedures.

1. Construction of the list {h} of all possible helices for a given RNA sequence.

2. Generation of the initial population of RNA secondary structures (individuals) comprising N structures. Each secondary structure consists of a particular subset of helices taken from {h}.

3. Calculation of the energy for every RNA secondary structure in the population considered.

* http://wwwmgs.bionet.nsc.ru/mgs/programs/2dstructrna.

** The sequences used in the work can be provided on request.

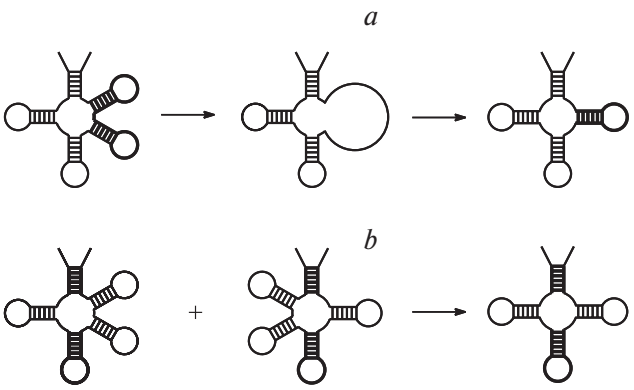

Fig. 2. Illustration of the operation of genetic operators in the GArna algorithm: (*a*) mutations and (*b*) recombinations. The mutating helices (*a*) and the backbone for the offspring structure (*b*) are marked by thick lines.

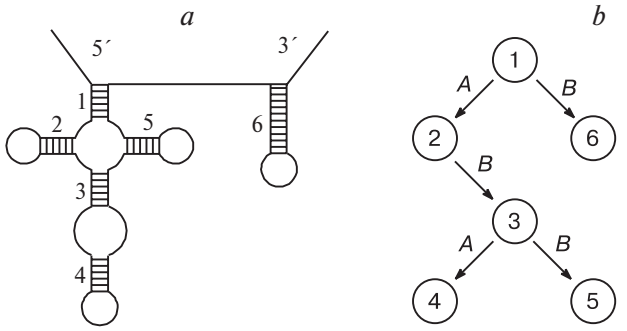

Fig. 3. Secondary structure of RNA (*a*) and its representation as a binary tree (*b*). Representation (*b*) is similar to secondary structure representations proposed previously.[22,23]

4. Decrease in the population down to a specified level through selection.

5. Carrying out recombinations between pairs of selected individuals and filling the vacancies resulting from step 3 by "progenies".

6. Conduction of multiple mutations (local structure changes).

7. Return to step 3 until either of two conditions is met:

— a specified population degeneracy has been attained, *i.e.*, a sufficient number of similar secondary structures (relatives) exists;

— a specified number of optimization cycles (evolution generations) has been carried out.

8. Selection of the minimum-energy structure from the last generation as the result of calculations.

Now we proceed to the description of the algorithm steps with more detailed consideration of its features.

The construction of the set {h} of all possible helices for the considered RNA molecule includes also the helices obtained from the initial one by removal of one or several successive complementary pairs. The minimum length of the helix is specified by the user and can be equal to one or more complementary pairs. The minimum length of the hairpin loop equals three nucleotides.

The initial population. In the procedure of definition of the initial state, we set ourselves the task to create a set of RNA secondary structures differing most appreciably from one another in order to avoid the premature convergence of the algorithm. The initial population size is taken to include 100 individuals, and it is maintained constant during the evolution. The initial structures were formed by adding helices from {h} according to standard stereochemical compatibility rules. This list was divided into two sublists, $\{h\}_1$ comprising helices that have already been used in the formation of the initial population and $\{h\}_2$ comprising the other helices. At the first stage, each initial structure was formed, first of all, from random helices taken from $\{h\}_2$, which thus moved to $\{h\}_1$. After the {h} list was exhausted, the structures were constructed from helices of the $\{h\}_1$ list.

Calculation of the energy of RNA secondary structure. The energy of RNA secondary structure was calculated using a fast recursive procedure. A structure was represented as a binary tree (Fig. 3). The tree nodes were matched by helices and the edges connecting the nodes were matched by all loops except for hairpin loops. The first helix from the sequence 5′-end is the tree root (Fig. 3, helix *1*), whereas the terminal points of the tree correspond to hairpins. Each node has two pointers, pointer *A* to the subtree (substructure), which is closed by the helix corresponding to the node, and pointer *B* to the substructure located at the 3′ position relative to this helix.

The structure energy is calculated by an ordered tree traversal until all nodes are visited. The sequence of steps from each node is as follows: first, along pointer *A*, then along pointer *B* and, finally, back one level up. This procedure is called "top-down" traversal and allows simple recursion.[21] In the example shown in Fig. 3, the points will be visited in the order 123456. Upon each move, the energy of the helix corresponding to the visited node and the energy of the traversed loop (with account for all helices that form this loop) are added to the structure energy. The thermodynamic parameters from the compilation were used.[15]

Selection. At each stage of the selection, the decision concerning the survival of a particular individual was based on a stochastic procedure in which the survival probability was determined by the difference between the fitness of an individual and the population mean value. The fitness of an individual (secondary structure) was calculated as follows:

$$f_i = \exp(-E_i/\Delta E),$$

where $E_i < 0$ is the free energy of structure i; $\Delta E > 0$ is the effective energy resolution, *i.e.*, the difference between the energies of structures such that the ratio of their fitnesses is equal to e. The calculations showed that

the optimal ΔE = 3 kcal mol^{-1}. A decrease in ΔE with respect to the optimal value reduces the rate of convergence of the algorithm. When ΔE increases, the algorithm converges to shallower minima.

Mutations. In our algorithm, mutations represent a controlled process and resemble the adaptive search procedure.[24] In GA, such an implementation of mutations has been used previously to calculate the hairpin conformations at the atomic level.[7] Mutations are fixed only if a specified neighborhood of the secondary structure considered contains another, more stable structures.

In our algorithm, mutations comprised several steps (Fig. 2):

— a specified number of individuals (structures) for mutation was randomly selected out of the population;

— a specified number of helices was eliminated from the structure;

— helices most favorable as regards the structure stability were successively added.

If the resulting structure was inferior in energy to the initial structure, then the mutation outcome was rejected. The mutation stage resembles some known algorithms for simulation of the secondary structure formation.[19,25,26]

Recombinations. When a large-scale search is carried out, recombinations represent a unique point that distinguishes GA from other stochastic optimization algorithms. In our algorithm, recombinations are directed at equal and, hence, the greatest distinction of the progeny from both parents and thus ensure the largest-scale search. Two randomly chosen parent structures and their common helices formed the backbone of the progeny structure. Then the backbone was completed by alternating addition of random helices from the parent structures (Fig. 2, *b*). When all the parent helices became incompatible with those already formed, the structure was completed by helices from the total list {h}.

Termination of the calculations. Testing of GArna with the N and ΔE parameters mentioned above showed that degeneration of the population (*i.e.*, similarity of the secondary structures forming the population) indicates that optimization is situated in the region of a global or a deep local minimum and will not leave it in the future. In this situation, further calculations result only in accumulation of replicas of the optimal structure, and, therefore, the calculation can be terminated at this point. The similarity of two structures can be conveniently evaluated by the number of coinciding helices. This number, normalized and averaged over all pairs in the population, characterizes the population degeneracy

$$D = \frac{2}{N(N-1)\max\limits_i K_i} \sum_{i=1}^{N} \sum_{j=i+1}^{N} k_{ij}$$

where K_i is the number of helices in structure i, k_{ij} is the number of helices common to structures i and j; N is the population size. The non-negative parameter D equals 1 if and only if the whole population is represented by replicas of a single individual. During evolution, the D parameter increases and tends to 1, the calculations being terminated when the D value exceeds a specified threshold D_c. In our calculation of the global minimum (in this work), D_c was taken to be 0.9. For smaller D_c values, the algorithm produces the most stable structure for the moment where $D = D_c$; this feature can be used for the calculation of metastable secondary structures.

Results and Discussion

Owing to its high running speed, the genetic algorithm described above is suitable for mass calculations in the analysis of secondary structures of RNA. Examples of such calculations are given below.

1. Dependence of the energy of secondary structures of random RNA sequences on the nucleotide composition

The development of methods for fast assessment of stability of the secondary structures of long RNA sequences without direct calculation of the secondary structure is important for practical purposes, for example, for the search for RNA in genome sequences. One possible approach is based on introduction of an index that reflects the complementarity potential and can be estimated from the contextual features of the RNA nucleotide sequence; in the simplest case, from nucleotide frequencies.

As will be seen from the results, two circumstances should be taken into account in constructing such an index. First, the mere account of the G+C content of an RNA sequence is insufficient because A—U and G—U couples also contribute to the formation of the secondary structure. Second, the ratio of the frequencies of the nucleotides able to enter into complementary interactions is also important. This is especially significant when the frequencies of these nucleotides are appreciably different. In the deficiency of any of them, the complementary nucleotide would not have a partner to bind with.

To study this point, we carried out the following calculations using the GArna algorithm. We considered random RNA sequences, 50 and 250 nt long, with all possible nucleotide frequencies (n_g, n_c, n_a, n_u) divisible by 0.1 that satisfy the normalization condition, $n_g + n_c + n_a + n_u = 1$. The sequences having no complementary nucleotides were excluded from consideration, altogether 255 combinations of such nucleotide frequencies were analyzed. For each combination, 100 sequences with a random order of nucleotides were generated. Thus, altogether $2 \cdot 255 \cdot 100 = 51000$ random sequences were stud-

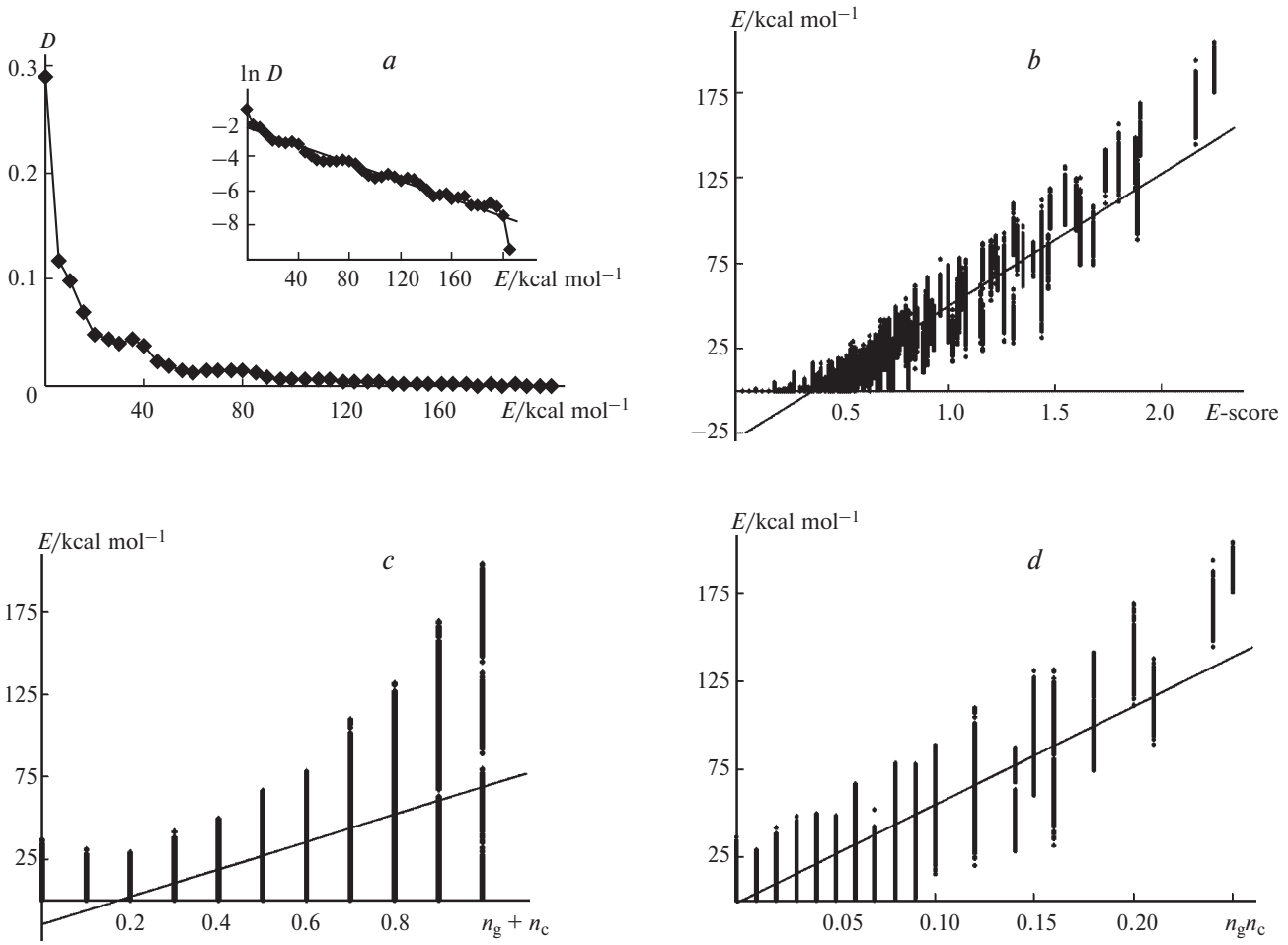

Fig. 4. Statistics of the secondary structure energies for random sequences of variable composition: energy distribution (the inset shows linearization in the semilogarithmic coordinates) (*a*) and energy (*E*) *vs.* complementarity index (*E*-score) (*b*) the $n_g + n_c$ sum (*c*), and the $n_g n_c$ product (*d*) (*D* is the fraction of structures with the energy *E*).

ied. For each of them, a low-energy secondary structure was calculated using the genetic algorithm.

The energy of these structures was found to vary over a broad range (Fig. 4, *a*, the energy is presented with an opposite sign). It is of interest that the number of sequences with equal energies *E* decreases exponentially as a function of *E*. This means that the fraction of sequences able to form a stable secondary structure is very low.

We used several indices for quantitative characterization of the secondary structure stability in terms of the nucleotide context. It was found that the simplest index, $n_g + n_c$, is slightly connected with the structure energy ($r^2 = 0.37$, Fig. 4, *c*); this sum does not take into account the unbalance of G and C nucleotides, *i.e.*, the fact that a nucleotide needs a complementary partner to become coupled within a secondary structure. Therefore, a better correlation was found for the product $n_g n_c$ ($r^2 = 0.82$, Fig. 4, *d*). Further we made this index more general by including the A—U and G—U pairs into consideration and introduced a complementarity index called *E*-score:

$$E\text{-score} = 9n_g n_c + 3n_a n_u + 2n_g n_u, \quad (1),$$

where $n_i < 1$ is the frequency of nucleotide *i* in the sequence. The factors appearing at the pairs of frequencies for complementary nucleotides, G—C, A—U, and G—U reflect their energy contribution to the secondary structure formation. It can be seen (see Fig. 4, *b*) that the *E*-score is well correlated with the secondary structure energy E_{rand} ($r^2 = 0.89$). The corresponding linear regression equation

$$E_{rand} = -77.62E\text{-score} + 27.96$$

can be used for fast estimation of the contribution of the nucleotide composition of RNA molecules to the secondary structure energy, in particular, for the search for

noncoding RNA in the genome sequences. The relation can be easily extended to RNA with arbitrary lengths (see Section 2) or to the interaction of two RNA molecules.

2. Study of the energy distribution of the secondary structures of random sequences with different nucleotide frequencies

Let us proceed to the analysis of the energy fluctuations for the optimal secondary structures in random sequences having fixed nucleotide frequencies and lengths. In this case, energy variation reflects the effects of nucleotide arrangement.

2.1. The first and second moments of distribution. For random sequences and the series of sets of nucleotide frequencies $\{n_i\} = (n_g, n_c, n_a, n_u)$, the energy of the optimal secondary structure averaged over the nucleotide arrangement was shown[24] to increase linearly with the sequence length L. Evidently, this feature is retained for any fixed nucleotide composition $\{n_i\}$. In other words, nucleotide frequencies $\{n_i\}$ determine the proportionality factor in this linear relation

$$\langle E_i \rangle = E_0(\{n_i\})L. \qquad (2)$$

Here, we omitted the L-independent summand,[24] significant only for small L (Fig. 5). The standard deviation of energy $D = \sqrt{\mathrm{var}(E_i)}$ also follows a linear dependence[24] (Fig. 5)

$$D_i = D_0(\{n_i\})L. \qquad (3)$$

The results of this section are important for substantiating the applicability of standard statistical tests to the analysis of the secondary structures of natural RNA and will be used below. In addition, Eqs. (2) and (3) provide a simple performance criterion for GA optimization, the deviation from linearity at great L being indicative of capture of the algorithm by a local minimum. It follows from Fig. 5 that our algorithm finds the optimal secondary structure for sequences up to 300 nt long.

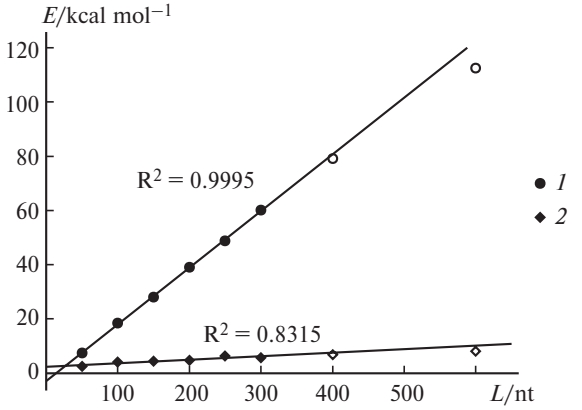

Fig. 5. Mean value (*1*) and standard deviation (*2*) of the secondary structure energy (E) *vs.* length (L/nt) of a random sequence of an equal composition. Each point was obtained by averaging over five sequences. The regression line is drawn through filled characters (the presumptive region of GA convergence to the optimal structure). The energy variance fluctuates due to the small size of sampling.

2.2. Distribution pattern of E_i. By nature, the $\{E_i\}$ distributions belong to the type of limiting distributions; however, previously, they have been noted[24,27] to bear a visual resemblance to normal distribution. The answer to the question as to whether approximation of $\{E_i\}$ by a normal distribution is justified is important for the use of standard statistical tests in the analysis of the secondary structures for natural sequences. To evaluate the similarity of $\{E_i\}$ and a normal distribution, the following numerical experiment was carried out.

Sequences with a length of 250 nt were isolated out of the sampling used in Section 1. In all, there were 255 groups of random sequences, 100 sequences in each group, with identical contents of nucleotides. Of 255 groups, 175 groups were selected in which all sequences had a secondary structure with a negative energy. These 175 groups were subjected to the Shapiro—Wilks W-test from the Statistica 5.0 package* to check the similarity to a normal distribution. An example of calculations is presented in Fig. 6, *b*. The energy distributions for 152 groups of the 175 groups proved to be indistinguishable from the normal distribution with the significance level $p = 0.05$. The distinction of the other 23 distributions from the normal distribution might be due to statistical fluctuations caused by the small number of sequences in the sampling.

2.3. Relative deviation technique. In the previous section, it was shown that $\{E_i\}$ is approximated adequately by a normal distribution. Therefore, the value

$$Z_i(L = \mathrm{const}) = \frac{E_i - \langle E_i \rangle}{\sqrt{disp(E)}} \qquad (4)$$

obeys, to the same accuracy, the standard normal distribution, *i.e.*, the normal distribution with a zero mean and unit variance.[27] With allowance for Eqs. (2) and (3), this implies that the value

$$Z_i = \frac{\frac{E_i}{L} - E_0(\{n_i\})}{D_0(\{n_i\})} \qquad (5)$$

also obeys the standard normal distribution, but in this case, for a sampling of sequences of variable length and composition.

Here the distribution parameters E_0 and D_0 depend on the nucleotide frequencies $\{n_i\}$. To calculate this de-

* http://www.statsoft.ru

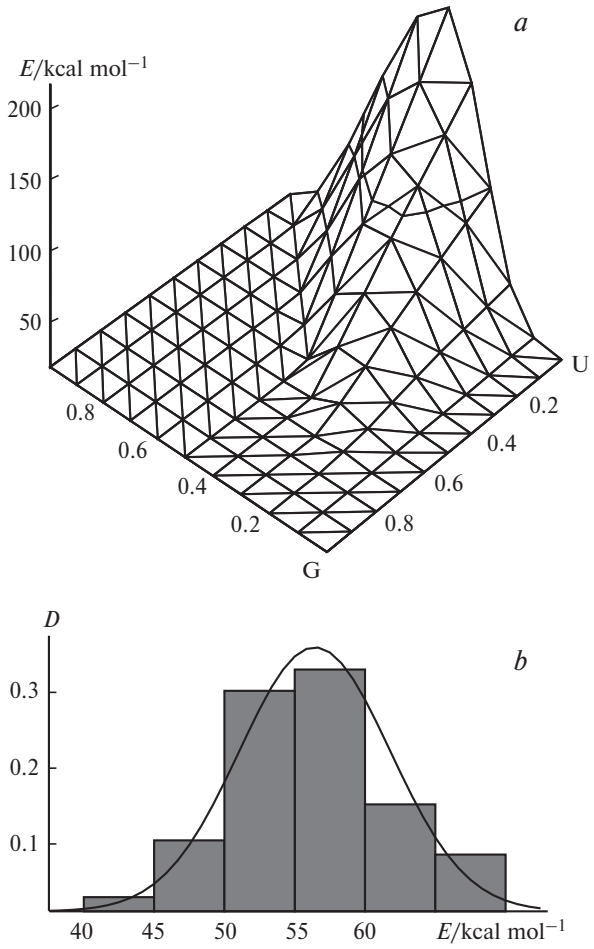

Fig. 6. Examples of secondary structure energy (E) distribution for random sequences with a length of 250 nt: (*a*) $<E>$ *vs.* n_u and n_g ($n_a = 0$); (*b*) secondary structure energy distribution for 100 random sequences with the frequencies $n_a = 0.3$, $n_u = 0$, $n_g = 0.4$, and $n_c = 0.3$ and its approximation by a normal distribution (continuous line).

pendence, we used 255 E_0 and D_0 values calculated above in the assessment of the $\{E_i\}$ similarity to a normal distribution. For intermediate $\{n_i\}$ values, we used a linear interpolation.

It was found that (to an accuracy of a 10% step in the nucleotide composition) both the $E_0(\{n_i\})$ and $D_0(\{n_i\})$ functions increase monotonically to a single maximum in the region of GC-rich sequences (Fig. 6, *a*). The vicinity of this maximum corresponds to the distribution tail area in Fig. 4, *a*. With allowance for Eqs. (2) and (3), this finding indicates that, as the sequence becomes longer, the energy surface in the space of sequences $\{i\}$ becomes more and more irregular. This irregularity is most pronounced in the $\{i\}$ area that contains sequences with the most stable secondary structures.

Now we consider a sampling of real sequences and calculate Z_i for each sequence i of this sampling. It is significant that the sampling may be heterogeneous regarding RNA lengths and compositions. Recall that Z_i for random sequences (null hypothesis) obey a standard normal distribution. A positive mean $<Z>$ value in a real sampling implies a low potential of the nucleotide context of natural RNA toward the formation of a stable secondary structure with respect to random sequences of the same composition. Conversely, if $<Z> < 0$, the sequences of the group under study have an enhanced potential toward the formation of the secondary structure, which is lost on thorough mixing of nucleotides.[27—29]

Table 1 illustrates this conclusion for two RNA classes that form stable secondary structures, tRNA and 5S rRNA. Both classes are characterized by reliably negative $<Z>$ values. This is consistent with the observed difference between tRNA and random sequences.[30] A more precise value, $<Z> = -1.65$, was obtained for tRNA quite recently by mass calculations for the set[27] containing 1451 sequences.

In section 3, the relative deviation technique (Z-score) is used for the analysis of eukaryote mRNA 5′-UTRs.

3. Analysis of the secondary structure of the mRNA 5′-UTR for plants and mammals

The translation of the eukaryote mRNA is initiated by a scanning mechanism.[31] The ribosome 40S-subunit recognizes the cap and binds to the 5′-end of mRNA, in order to migrate along the molecule up to the initiating codon. When the 40S subunit reaches this codon, subunit 60S joins it to start the translation. The mRNA hairpin structures can slow down the migration of subunit 40S both by themselves and as complexes with translation-inhibiting proteins.

It is worthy of note that stability of the secondary structures of the 100 mRNA nucleotides closest to the cap site is the most efficient characteristics for gene classification in terms of the expression level.[18] However, this effect can be explained by mere separation of groups of sequences in terms of the nucleotide contents, which has also been observed previously.[16] To investigate this point, we analyzed the 5′-UTR for plants and mammals (for the sampling formation, see System and methods) and calculated the Z-score and E-score values mentioned above. Recall that the E-score reflects the contribution of the nucleotide composition to the secondary structure stability (composition effect). The Z-score evaluates the untypicalness of the secondary structure stability (*i.e.*, the E value is taken as a null hypothesis, reflecting the ordering in the nucleotide arrangement).

3.1. Dicot plants. Generally, the 5′-UTRs of dicot plants are characterized by negative Z values (see Table 1). This may be indicative of the fact that these sequences have acquired an additional capacity for the formation of a secondary structure during evolution (owing to the

Table 1. Characteristics of the primary and secondary structures of natural sequences

RNA	Expression level	Number	G+C (%)	$\langle E \rangle$	$\langle Z \rangle$
tRNA		14	58.9±2.8	1.03±0.06	$-1.84\pm0.71^{+}$
5S-rRNA		7	55.1±1.8	96.0±0.04	$-1.81\pm1.38^{+}$
mRNA 5′-UTR					
human	H	51	58.8±11.2	0.97±0.22*	$-0.48\pm1.38^{+}$
	L	202	61.3±12.6	1.09±0.25*	$-0.69\pm1.25^{+}$
	All	253	60.8±12.3	1.06±0.24	$-0.65\pm1.28^{+}$
dicot plants	H	44	37.1±7.7	0.66±0.12	$-0.19\pm1.22^{*+}$
	L	22	35.9±6.7	0.64±0.15	$-0.9\pm1.43^{*+}$
	All	66	36.7±7.3	0.64±0.13	$-0.43\pm1.33^{*+}$
mammals	H	33	53.7±12.5*	0.93±0.21*	0.26±1.1*
	L	17	62.3±11.7*	1.07±0.26*	−0.26±1.57*
	All	50	56.6±12.8	0.96±0.24	0.08±1.29

Note. The indices show the statistically significant differences for the significance level $p < 0.05$, either appearing as a result of selection of random sequences ("+") or reflecting the difference between H- and L-mRNA ("*").

local ordering of the nucleotide context). However, judging by the average value $\langle Z \rangle = -0.43$ (see Table 1), these secondary structures are not as stable as the tRNAs and 5S-rRNAs we studied, which have $\langle Z \rangle = -1.84$ and −1.81, respectively.

The group of dicot mRNA considered was heterogeneous and contained mRNA of both high (H) and low (L) expression genes. In this connection, it is of interest to compare the $\langle Z \rangle$ values for the 5′-UTRs of these two mRNA subgroups. It was found (see Table 1) that taken separately, these groups follow the general regularity with negative $\langle Z \rangle$ values. Meanwhile, the $\langle Z \rangle$ value for the L-group is 4.7 times greater in magnitude than that for the H-group. This means that the 5′-UTRs of low-expression mRNA of dicot plants can form more stable secondary structures than the 5′-UTRs of high-expression gene mRNA.

Note that the *Z*-score proved to provide much more information for this group of sequences than the G+C composition, which shows almost no difference between H- and L-RNA (see Table 1). The close G+C content values for H- and L-RNA of plants are in line with the lack of difference between these RNAs regarding the *E*-score (see Table 1).

Thus, it was found that the 5′-UTRs of H- and L-mRNA are similar in the nucleotide contents and differ in the stability of the secondary structure. This may imply that the distinctions in the stability of the secondary structures of the regions arisen during evolution were due to nucleotide ordering, which allows the formation of stable secondary structures, rather than due to the nucleotide composition divergence. This evolutional editing occurred mainly in L-RNAs. In other words, of all the 5′-UTR of plants and mammals we studied (see below), it is in this group that evolutionary conservative signals based on RNA secondary structure and repressing the rate of translation initiation could be found with the highest probability.

3.2. Mammals. The $\langle Z \rangle$ value for the 5′-UTRs of mammals proved to be close to zero (see Table 1). When considering separately the 5′-UTRs of H- and L-mRNAs of mammals, we found that the $\langle Z \rangle$ value for L-RNAs is slightly negative, whereas for H-RNAs, this value is slightly positive. For the sampling sizes used, both deviations of $\langle Z \rangle$ from zero are insignificant, however the differences between H- and L-RNAs are reliable. The lower *Z*-score value for L-RNAs coincides qualitatively with the results of analysis of human 5′-UTRs (see Table 1) and with the data described above for the mRNA 5′-UTRs of dicot plants. However, the human H- and L- mRNAs are both characterized by negative $\langle Z \rangle$, which is typical of plants but not of mammals. This contradiction might have arisen due to the different principles of formation of the human and mammal 5′-UTR samplings, in particular, due to the small size of the latter. This sampling should be increased in order to answer the question of whether the local ordering of the nucleotide context resulting in an enhanced stability of the 5′-UTR secondary structures is a general feature for the studied classes.

Nevertheless, both for mammals and humans, L-RNA contains, on average, more G and C nucleotides and has greater *E*-score index than H-RNA (see Table 1). Thus it can be concluded that the secondary structure of L-RNA is more stable than that of H-RNA. The results concerning the relationship between the expression level of mammal and human genes and the energy of the 5′-UTR secondary structures of the corresponding mRNAs are in qualitative agreement with both the results of the above-described analysis of plant mRNAs and published data.[18]

Generally, the mRNA 5′-UTRs of plants and mammals differ appreciably in the average contents of nucle-

otides G and C (see Table 1), plant 5′-UTRs being depleted, while the 5′-UTRs of mammals being enriched in these nucleotides. This is consistent with the known fact of higher contents of G+C in the genomes of warm-blooded organisms compared to plants.[32] In addition, the plant 5′-UTRs studied have almost identical G+C compositions, whereas in the case of mammals, L-RNAs are richer in nucleotides G and C than H-RNAs. The genomes of mammals are known to have an isochoric genome organization, which is manifested as the presence of extended regions with relatively homogeneous nucleotide composition.[33] This may imply that the high- and low-expression genes of mammals, which differ appreciably in the G+C-composition, are located in different isochores.[34] Quite recently, the assumption stating that the expression of the genes of mammals is related to their isochore localization has been confirmed experimentally.[35]

* * *

Thus, we have developed the GArna program package designed for the analysis of RNA secondary structures. The central item in the package is the fast GArna genetic algorithm for the search of low-energy RNA structures. Although GArna operates more slowly than dynamic algorithms, it has the same time complexity (dependence of the running time on the sequence length, Fig. 7, *a*) sufficient for the analysis of the RNA secondary structure through the Internet and for mass calculations on a PC. An advantage of the algorithm is that it provides the possibility of predicting metastable RNA structures by two methods, either by terminating the calculations before the convergence of the algorithm or by decreasing the selection force (see the section Algorithm, clauses 4 and 7, respectively). Besides the options listed above, the GArna software can be used for the analysis of RNA secondary structures taking into account antisense interactions.

The high running speed of the algorithm is attained due to some its specific features:

(1) a special method for setting the initial state (quasi-uniform coverage of the space of secondary structures by the initial population);

(2) adaptive mutation process depending on the local features of the energy surface of a secondary structure and resulting in acceleration of movement between separated minima;

(3) symmetrization of recombinations directed at the maximum distinction of the offsprings from the parents, which ensures extensive probing of the space of potential secondary structures;

(4) fast recursive procedure for the calculation of the secondary structure energy.

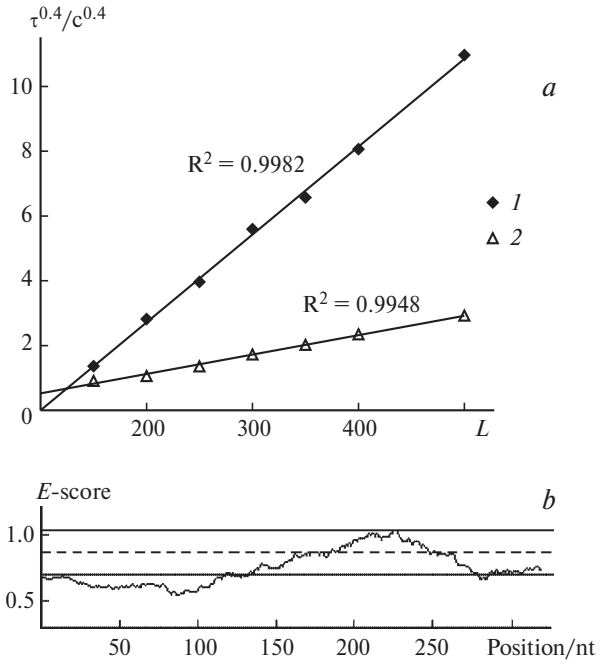

Fig. 7. *a*. Typical dependence of the running time (τ) on the sequence length (*L*) for the GA (*1*) and mfold algorithm (*2*) (averaging over 50 random sequences of equal composition). *b*. The *E*-score profile constructed with a window of 80 nt for sequences from the EMBL bank (length 400 nt, the tRNA Ser occurs between positions 218 and 299).

The genetic algorithm was used to study the mRNA 5′-UTRs of high- and low-expression genes of plants and mammals. These groups of genes were found to differ from each other in several contextual and structural characteristics (see Table 1).

Using GArna, mass calculation of the secondary structures of random RNAs was carried out and two methods for evaluating the secondary structure stability were analyzed. One method is based on the estimation of energy typical of sequences with nucleotide content identical to the sequence under study. This index, *E*-score, can be used for analysis of a large number sequences and long RNA, because in these cases, direct calculation of the secondary structure using the fastest algorithms requires prolonged calculations. However, the attempts to use this index for the search for noncoding RNA in genome sequences were, generally, unsuccessful (one of the sparse successful examples is shown in Fig. 7, *b*).

The second parameter, *Z*-score, provides additional information on the RNA by determining the extent to which its sequence its nontypical concerning the ability to form a stable secondary structure. The method is based on the comparison the RNA stability with the stability of the sequences obtained from the RNA by multiple random permutations of nucleotides. This approach was first used within a window with a fixed length sliding along

the sequence.[28] Recently, this procedure was used to perform mass analysis of mRNAs[29] and to study the problem of RNA search in genome sequences[27] with pessimistic conclusions. In this work, we substantiated the approximation of the distribution of Z by a normal distribution; this made possible the use of standard statistical tests for analysis of natural sequences, inhomogeneous in length and nucleotide composition. We calculated parameters of this distribution, depending on the nucleotide composition and the RNA length, which markedly accelerated the calculation of Z-score and allowed performing this analysis through the Internet.

Due to the small size and weak homology of noncoding RNA, it remains obscure how many of them have not yet been identified in the sequences that have already been sequenced. The computer algorithms for the solution of "ribonomics" tasks,[36] *i.e.*, for the search for and determination of the biological functions of new RNAs are at an early stage of development. Taken separately, the E-score and Z-score cannot be recommended for the search for new RNAs. The joint use of these values could be more efficient, in particular, in combination with comparative analysis procedures and the search for structural patterns (S- and U-turns, tetraloops, *etc*.).

This work was financially supported by the Russian Academy of Sciences (the Integration Project of the Siberian Branch of the Russian Academy of Sciences, No. 65), the Russian Foundation for Basic Research (Projects No. 98-07-91078, No. 98-07-90126, No. 99-07-90203, No. 99-04-49879, and No. 00-07-90337) and the National Program "Human Genome".

References

1. N. A. Kolchanov, I. I. Titov, I. E. Vlassova, and V. V. Vlassov, *Prog. Nucl. Acids Res. Mol. Biol.*, 1996, **53**, 196.
2. M. Zuker, *Curr. Opin. Struct. Biol.*, 2000, **10**, 303.
3. D. F. Mathews, J. Sabina, M. Zuker, and D. H. Turner, *J. Mol. Biol.*, 1999, **288**, 911.
4. J. McCaskill, *Biopolymers*, 1990, **29**, 1105.
5. A. A. Mironov and A. E. Kister, *J. Biomol. Struct. Dyn.*, 1986, **4**, 1.
6. A. Fernandez, *Phys. Rev. (E)*, 1993, **48**, 3107.
7. H. Ogata, Y. Akiyama, and M. Kanehisa, *Nucl. Acids Res.*, 1995, **23**, 419.
8. A. P. Gultyaev, F. H. D. van Batenburg, and C. W. A. Pleij, *J. Mol. Biol.*, 1995, **250**, 37.
9. G. Benedetti and S. Morosetti, *Biophys. Chem.*, 1995, **55**, 253.
10. K. M. Currey and B. A. Shapiro, *Comput. Applic. Biosci.*, 1997, **13**, 1.
11. V. Proutski, E. A. Gould, and E. C. Holmes, *Nucl. Acids Res.*, 1997, **25**, 1194.
12. I. I. Titov, V. A. Ivanisenko, and N. A. Kolchanov, *Comput. Techn.*, 2000, **5**, 48.
13. S. Forrest, *Science*, 1993, **261**, 872.
14. D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, San Mateo, CA, 1989.
15. J. A. Jaeger, D. H. Turner, and M. Zuker, *Proc. Natl. Acad. Sci. USA*, 1989, **86**, 7706.
16. A. V. Kochetov, M. P. Ponomarenko, A. S. Frolov, L. L. Kisselev, and N. A. Kolchanov, *Bioinformatics*, 1999, **15**, 704.
17. E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Pruss, I. Reuter, and F. Schacherer, *Nucl. Acids Res.*, 2000, **28**, 316.
18. R. V. Davuluri, Y. Suzuki, S. Sugano, and M. Q. Zhang, *Genome Res.*, 2000, **10**, 1807.
19. J. P. Abrahams, M. van den Berg, E. van Batenburg, and C. W. A. Pleij, *Nucl. Acids Res.*, 1990, **18**, 3035.
20. W. Fontana, D. A. M. Konnings, P. F. Stadler, and P. Schuster, *Biopolymers*, 1993, **33**, 1389.
21. N. Wirth, *Algorithms and Data Structure*, New Jersey 07632, Prentice-Hall, Inc., Englewood Cliffs, 1986.
22. B. A. Shapiro, *Comput. Applic. Biosci.*, 1988, **4**, 387.
23. M. Zuker and D. Sankoff, *Bull. Math. Biol.*, 1984, **46**, 591.
24. W. Fontana, T. Griesmacher, W. Schnabl, P. F. Stadler, and P. Schuster, *Monatsh. Chemie*, 1991, **122**, 795.
25. H. M. Martinez, *Nucl. Acids Res.*, 1984, **12**, 323.
26. R. Nussinov and G. Pieczenik, *J. Theor. Biol.*, 1984, **106**, 244.
27. E. Rivas and S. R. Eddy, *Bioinformatics*, 2000, **16**, 583.
28. S. Y. Le and J. V. Maizel, *Nucl. Acids Res.*, 1997, **25**, 362.
29. W. Seffens and D. Digby, *Nucl. Acids Res.*, 1999, **27**, 1578.
30. P. G. Higgs, *J. Phys. I. France*, 1993, **3**, 43.
31. M. Kozak, *Biochimie*, 1994, **76**, 815.
32. G. Bernardi and G. Bernardi, *J. Mol. Evol.*, 1986, **24**, 1.
33. G. Bernardi, *Gene*, 2000, **241**, 3.
34. I. I. Titov, D. G. Vorobiev, and N. A. Kolchanov, *Intern. Conf. on Bioinformatics of Genome Regulation and Structure BGRS-2000* (*Novosibirsk, August 7—11, 2000*), Novosibirsk, 2000, 138.
35. H. Caron, B. van Schaik, M. van der Mee, F. Baas, G. Riggins, P. van Sluis, M.-C. Hermus, R. van Asperen, K. Boon, P. A. Voute, S. Heisterkamp, A. van Kampen, and R. Versteeg, *Science*, 2001, **291**, 1289.
36. V. Bourdeau, G. Ferbeyre, M. Pageau, B. Paquin, and R. Cedergren, *Nucleic Acids Res.*, 1999, **27**, 4457.

Received October 12, 2001;
in revised form April 16, 2002