ISSN 0006-3509, Biophysics, 2017, Vol. 62, No. 6, pp. 857–864. © Pleiades Publishing, Inc., 2017.
Original Russian Text © D.A. Grechishnikova, M.S. Poptsova, 2017, published in Biofizika, 2017, Vol. 62, No. 6, pp. 1045–1053.

MOLECULAR BIOPHYSICS

The Physical and Geometric Properties of Human Transposon Stem–Loop Structures under Natural Selection[1]

D. A. Grechishnikova[a], * and M. S. Poptsova[b], **

[a] *Department of Physics, Moscow State University, Moscow, 119991 Russia*
[b] *National Research University Higher School of Economics, Moscow, 101000 Russia*
**e-mail: daria.grechishnikova@gmail.com*
***e-mail: maria.poptsova@gmail.com*

Received July 11, 2017

Abstract—Secondary RNA structures play an important role in transposition, in particular, in RNA recognition by transposon proteins. Previously, we found a conserved structure at the 3'-end of human transposons and proposed a hypothesis about the role of this structure in transposition. Although there is no similarity at the sequence level, the conserved position of this structure points to the fact that structural properties occur that are under positive natural selection. In this paper, the physical and geometric properties of stem-loop structures at the 3'-end of human transposons are identified and compared with properties of the structures of other genome regions. Each stem-loop structure was characterized by a set of ten characteristics: the Gibbs free energy, enthalpy, entropy, hydrophilicity, Shift, Slide, Rise, Tilt, Roll, and Twist. A model has been built using machine-learning methods, which recognizes stem-loop structures according to their physical and geometric characteristics with 94% accuracy. The most important parameters in the recognition model are hydrophilicity, enthalpy, Rise, and Twist. These properties of transposon structure are thought to be under positive natural selection.

Keywords: transposon, stem-loop structure, dinucleotide characteristics, entropy, Gibbs free energy, machine learning

DOI: 10.1134/S0006350917060070

INTRODUCTION

Transposons are DNA fragments that are able to multiply and move in the genome. Transposons are contained in the genomes of all eukaryotes and occupy a significant portion of them: 46% of the human genome, 37% of the mouse genome, 10% of the fruit fly genome, and 85% of the maize genome [1, 2]. It was believed for many years that transposons are so-called "junk DNA" and do not have any functional significance. However, after the discovery of hemophilia in an adult patient caused by a transposon jump into the gene for a coagulation factor, an increasing number of works began to appear devoted to the study of the pathogenic role of transposons. A total of 96 human diseases caused by transposon jumping are known. Currently, the role of transposons in the emergence of cancer is being actively studied; methods of using LINE-1 transposons as cancer markers are being developed [1, 3]. It has been established that transposons make a significant contribution to the variability of genomes of different organs, such as the brain and immune system [4]. There is also speculation that transposons are a tool of evolution, since they are able to cause both large-scale restructuring (e.g., nonallelic recombination between two elements of two different chromosomes) and small changes in the genome (duplications, inversions, and deletions) [5]. Moreover, transposons are able to influence their own expression and the expression of nearby genes. The transcription level depends on the tissue type and is affected by external factors, for example, by stress in the brain of mammals [6].

Transposons can be divided according to the type of movement into those that "cut and insert" and those that "copy and insert." DNA is copied into RNA, a copy of DNA is then created by reverse transcription of RNA, which is inserted into the genome. Transposons of the second type are called retrotransposons. LINEs (long interspersed elements) and SINEs (short interspersed elements) are the most common classes of human retrotransposons. The human LINE retrotransposons occur in approximately 500000 copies and occupy 17% of the genome, while SINE occur in more than 1 million copies and occupy 11% of the genome.

[1] *Abbreviations*: LINE, long interspersed elements; SINE, short interspersed elements.

LINEs use their own molecular devices for transposition, as encoded to copy their sequence and insert it into the genome. SINES are parasitic: they do not encode proteins and they require LINE proteins for transposition. It remains unclear how a protein recognizes its own RNA and SINE RNA [6]. It has been shown for several organisms that the protein recognizes a secondary structure of a stem–loop type at the 3'-end of RNA. Moreover, it has been shown that some organisms have identical sequences at the 3'-end of the LINE and SINE transposons [7–9]. In such cases, it is assumed that LINE proteins "recognize" the secondary structure at the 3'-end, which is the same in the SINE and LINE RNA transcripts. In the cases where transposons do not have the same 3'-end, it is considered that the poly-A-tail, which is present in LINE and SINE transposons, is recognized. However, almost all mRNAs have poly-A-tails, which casts doubt on selective recognition based on binding with them. The mechanism of retrotransposition is not fully understood as yet. One of the most important issues is the manner of the LINE protein recognition of its own RNA transcript and of the SINE RNA transcript. We have previously detected a conserved secondary structure at the 3'-end of LINE and SINE transposon RNA in humans [10] and different species located throughout the tree of life (unpublished results). We hypothesize that this structure plays a role in the process of retrotransposition. Despite the complete lack of similarity at the level of sequences, the position of the conserved structure indicates the presence of properties that are under positive natural selection. It is most likely that structural features determine the protein binding with the stem–loop or hairpin structures.

The majority of proteins interact with a DNA molecule via contact with a region of 15–20 base pairs [11]. The local physico-chemical and geometric properties of the region play the most important role in interactions with protein. The dinucleotide level is the smallest approximation that makes sense to consider a sequence having physico-chemical and geometric properties. Recent studies have shown that the dinucleotide characteristics of RNA or DNA fragments can be used for prediction of protein binding [12, 13]. As an example, it has been shown that it is possible to build machine-learning models using physico-chemical and geometric characteristics of dinucleotides with high accuracy recognition of hot spots of recombination [14], splice sites [15], small regulatory RNAs resulting from transposon sequences [16], and the sites of DNA editing [13].

The aim of this work was to determine the physical and geometric properties of stem–loop structures at the 3'-end of human transposons and to compare them with the properties of structures from other regions of the genome. We chose dinucleotide properties available in the DiProDB database as the characteristics of the stem–loop structures in the human genome. Each stem–loop structure was characterized by the following set of characteristics: the Gibbs free energy, enthalpy, entropy, hydrophilicity, Shift, Slide, Rise, Tilt, Roll, and Twist. Using machine-learning methods we built a model that recognizes the structures of transposons according to their physical and geometric properties with 94% accuracy. The largest contributions to the recognition of structures were made by hydrophilicity, enthalpy, Rise, and Twist. It is suggested that these properties of the transposon structure are under positive natural selection.

MATERIALS AND METHODS

Genome annotation of secondary structure. Genome annotation of secondary structure was performed using DNA Punctuation software (www.dna-punctuation.org). The searching procedure for conserved secondary structures was described in detail in [10].

Stem–loop structure sampling. Four datasets have been formed: the conserved secondary structures from 6622 L1 human transposons and 39 of the most active L1 human transposons, the secondary structures taken from random genome locations, and randomly generated structures.

The physical and geometric characteristics of stem–loop structures. Thermodynamic characteristics have been calculated for each stem–loop structure on the basis of the nearest-neighbor model. In this model, the free energy of the structure is a sum of the free energy of the stem (paired parts) and of the loop (unpaired nucleotides). When the energy of the stem for each pair of nucleotides is calculated, the contributions from neighboring pairs on both sides are taken into account. Thus, the model suggests that the contribution of a base pair to a particular thermodynamic characteristic depends only on the two nearest neighbors. The thermodynamic characteristic are linearly dependent on the frequency of occurrence of a dinucleotide pair in a sequence. The free energy of an RNA duplex can be presented as

$$\Delta G^{37^\circ}_{\text{total}} = \sum_i n_i \Delta G^{37^\circ}(i) + \Delta G^{37^\circ}_{\text{initiation}} + \Delta G^{37^\circ}_{\text{sym}}.$$

The first term is the contribution of the i-th dinucleotide base pair that occurs n_i times in the sequence; i varies from 1 to 16 (the number of possible dinucleotide base pairs). The second term is the energy of initiation. It includes factors that are independent of the sequence (contrion condensation, the entropy loss upon duplex formation, etc.). The third term is responsible for entropy loss in the case of a duplex formed from a single DNA strand (complementary sites are on the same thread).

The dinucleotide physical characteristics of RNA were taken from the DiProDB database [17]. Ten characteristics were taken to build the model: the

Table 1. The thermodynamic characteristics of the stem–loop sequences

Data set	ΔG, kcal/mol	ΔH, kcal/mol	ΔS, kcal/mol/K
The structures of the active L1	−13.5 ± 0.9	−130.0 ± 2.1	−375.7 ± 5.2
The structures of the well-conserved L1 transposons	−11.1 ± 2.9	−119.0 ± 21.8	−347.8 ± 65.2
The structures from random genome regions	−11.9 ± 6.4	−120.7 ± 38.5	−350.7 ± 79.3
Randomly generated structures	−9.5 ± 3.0	−108.8 ± 26.7	−320.3 ± 77.5

nucleotide sequence of a secondary structure was split into dinucleotides, each of which was associated with a corresponding number of the DiProDB. The median was then computed. This procedure was conducted for each of the ten considered features. Thus, each sequence was mapped to a vector of ten numbers that correspond to ten physical characteristics–predictors. Two sets of the thermodynamic predictors were used: the average specific thermodynamic characteristics (per a dinucleotide pair) obtained as described above and the thermodynamic characteristics of the entire structure formation.

Four datasets were used for analysis: the sequences of the stem–loop type structures from currently active human genome L1 transposons, from the 6622 transposons with the best-conserved sequences, from the random regions of the genome, and the randomly generated sequences that form stem–loop structures.

Machine-learning model building. A machine-learning model was built using support vector machines for secondary structure recognition according to its physical and geometric characteristics. This method enables separation of the points in an n-dimensional space by an $(n - 1)$-dimensional hyperplane. A separating hyperplane is chosen from all possibilities, whose distance from the element of each class is at a maximum.

The following characteristics of the model were used to assess its performance: ACC, accuracy; SN, sensitivity; SP, specificity; and AUC, the ROC curve and the area under it.

The construction of spatial stem–loop structures. The 3DNA software package was used to build spatial structures [18]. It allows reconstruction of a structure from the dinucleotide characteristics of its sequence.

RESULTS

Six geometric characteristics were considered: three translational parameters and three angles that describe the spatial arrangement of one base pair relative to another. As well, we considered three thermodynamic characteristics: the Gibbs free energy, enthalpy, and entropy; as well as hydrophilicity.

Four datasets were used for the analysis: the sequences of the stem–loop type structures from currently active human L1 transposons, from the 6622 transposons with the best-conserved sequences, from the random regions of the genome, and randomly generated sequences that form a stem–loop structure. The thermodynamic characteristics were calculated for all of the sequences of the structures. The results are shown in Table 1. A tendency of all three parameters to achieve a minimum is observed for the active transposon structures.

The dependence of the average thermodynamic characteristics of the dinucleotide pair formation on the standard deviation for different classes is presented in Fig. 1. The graph shows that the random transposon structures can be separated according to their thermodynamic characteristics.

Many properties that are needed for protein binding can be described in terms of the geometric characteristics of dinucleotides, including the local elasticity, flexibility, and twisting. The average value and standard deviation of each of the six geometric characteristics were calculated for each sequence of the three data sets: the structures of the well-conserved L1 transposons, random regions of the genome, and randomly generated structures (Fig. 1). The structures of the most active and well-conserved L1 transposons were statistically significantly separated from the structures of the random regions and of the randomly generated genome according to all of the six geometric characteristics. The average values of the six geometric characteristics were calculated for each dataset by analogy with the thermodynamic characteristics and are given in Table 2. It is seen that the L1 transposon structures possess certain peculiarities.

Each stem–loop structure sequence has been mapped to a point in the ten-dimensional space of characteristics. A nine-dimensional hyperplane was constructed by the support vector machine, which separates the structures that belong to the end of the human L1 transposon from the structures from random locations of the genome, or randomly generated structures. The accuracy of the classifier is 94%.

Table 3 shows the parameters of the model. Figure 2 shows the ROC curve.

It is seen that the model is very effective. It enables one to determine the secondary structures that belongs to the 3'-end of the transposons in the human genome with high accuracy. The model is able to recognize the secondary structure in any given RNA sequence.

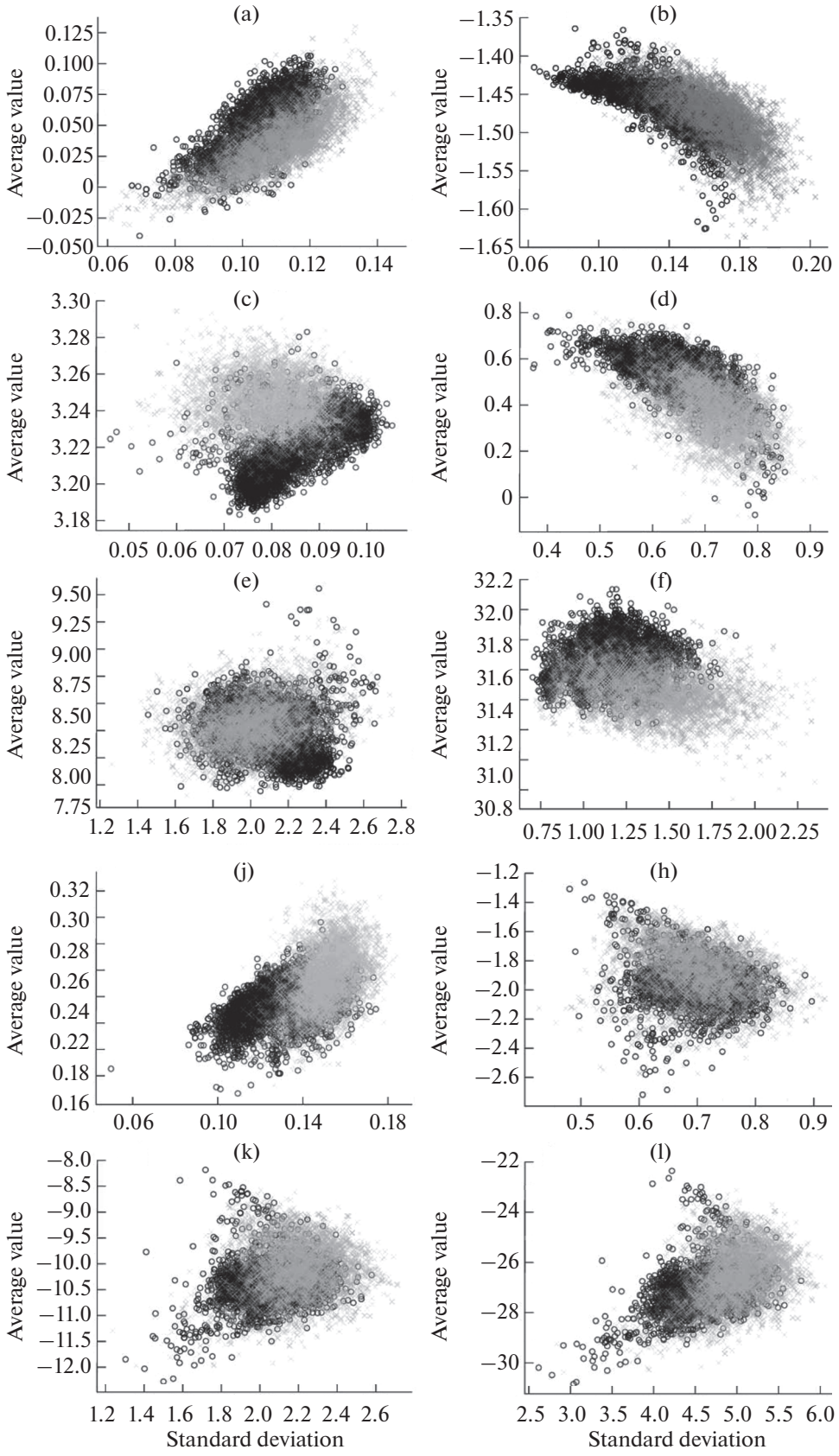
(a)
(b)
(c)
(d)
(e)
(f)
(j)
(h)
(k)
(l)
Average value
Standard deviation

Fig. 1. Comparison of the hairpins from the human L1 transposon 3'-end (in black) with the hairpins from the randomly generated sequences (in gray) according to the following characteristics: (a), the shift of a base pair relative to the adjacent pair in the direction of one of the grooves (Shift); (b), the shift of a base pair relative to the adjacent pair in the direction of the sugar phosphate backbone (Slide); (c), the helical pitch (Rise); (d), the opening angle of the neighboring base pairs in the direction of the sugar phosphate backbone (Tilt); (e), the opening angle of the neighboring base pairs in the direction of one of the grooves (Roll); (f), the angle of twist (Twist); (g), the hydrophilicity; (h), the Gibbs energy; (i), enthalpy; (j), entropy.

Since the hypothesis of a statistically significant difference between the characteristics of the transposon structures and the structures of the random genome regions was accepted, the next step of the analysis was to determine characteristics that differ between these two classes of structures. We used the Random forest machine-learning algorithm, which allows quantitative assessment of the importance (separating ability) of each independent variable (in our case, each characteristic). The results are presented in Table 4. The largest contribution to the recognition of structures was made by hydrophilicity, enthalpy, and the Rise and Twist parameters.

Model structures of the active L1 transposon and the random structure taken from the human genome were constructed based on dinucleotide geometric parameters (Fig. 3). The width of both grooves in the L1 transposon is smaller than in the random structure. Many proteins bind to sites located in the large groove. The smaller width in the case of the L1 transposon may be required for protein interactions [19]. The importance of the geometric characteristics of RNA for interaction with proteins has been experimentally shown for a large number of proteins that are involved in different processes of genome functioning, such as editing, processing, transport, and RNA interference [20, 21]. Further experimental study of the reverse transcriptase complex with the spatial structure of the transposon RNA is needed to determine the most important geometric characteristics of RNA for binding.

Some conclusions have been made regarding twists in these structures. The hairpin in the genome of the fish *Danio rerio* (Fig. 4a), which is recognized by reverse transcriptase, is twisted in a right-hand manner. Our analysis of the amino-acid sequences of the ORF2 35 protein from organisms from different levels of the evolutionary ladder revealed conserved domains in endonuclease and reverse transcriptase [10]. It is logical to assume that the similarity of the recognizing elements leads to the similarity of the elements that are to be recognized. We found the structure of the retrovirus reverse transcriptase complex with RNA in the PDB database (Fig. 4b). The general packing character indicates that overlapping α-helices tend to the formation of a left-hand superspiral.

Thus, the right-hand RNA hairpin interacts with a protein and recognizes a domain that has a tendency to the formation of a left-hand superspiral. This confirms the identified pattern: the interaction of the molecules of different types is carried out by structures that have different signs of their chirality [22].

DISCUSSION

The average values of the free energy, enthalpy, and entropy for structures from the currently active transposons are lower than for the structures from the well-

Table 2. The geometric characteristics of the stem–loop structure sequences and their hydrophilicity values

Data set	Shift, Å	Slide, Å	Rise, Å	Tilt, degr	Roll, degr	Twist, degr	Hydrophilicity
The structures of the active L1	0.07 ± 0.01	−1.45 ± 0.01	3.20 ± 0.01	0.6 ± 0.1	8.0 ± 0.1	31.9 ± 0.1	0.24 ± 0.01
The structures of the well-conserved L1 transposons	0.06 ± 0.02	−1.46 ± 0.03	3.22 ± 0.02	0.5 ± 0.1	8.1 ± 0.2	31.7 ± 0.1	0.23 ± 0.01
The structures from random genome regions	0.02 ± 0.03	−1.46 ± 0.06	3.24 ± 0.02	0.4 ± 0.2	8.3 ± 0.3	31.6 ± 2.5	0.24 ± 0.03
Randomly generated structures	0.04 ± 0.02	−1.48 ± 0.04	3.24 ± 0.02	0.4 ± 0.2	8.3 ± 0.2	31.4 ± 0.2	0.25 ± 0.22

Table 3. The parameters of the model

The model	Precision	Specificity	Sensitivity	Area under the ROC curve
SVM	0.94	0.94	0.93	0.98

Table 4. The separating abilities of the characteristics

The characteristics	Separating ability, %
The hydrophilicity	100.0
Enthalpy	45.3
Rise	30.0
Twist	28.2
Entropy	19.8
Shift	8.9
Slide	6.7

conserved L1 transposons, from random regions of the genome, and from randomly generated structures. Thus, the structures of the currently active L1 transposons are more stable than the other structures. The structures of the well-conserved L1 transposons are close on average to the structures from random regions of the genome. It is possible that stable and functionally important structures were in the latter sample, which increased its mean value. The randomly generated structures are the least stable.

The statistically significant difference between the geometric characteristics of the active well-conserved L1 transposon structures from the structures from random regions of the genome and the randomly generated structures indicates the preservation of geometric characteristics at the dinucleotide level in the process of evolution. Such conservation can be explained by maintenance of the structural features that are required for protein binding.

Many studies have shown that the RNA–protein binding complexes significantly depend on the RNA form and sequence (for a review see [23–25]). The size of the large and small grooves is associated with the Rise parameter. It is known that the size and configuration of the grooves play a significant role in the binding of proteins to RNA. The importance of the sizes of the major and minor grooves has been shown in adenosine deaminase binding to RNA in the process of RNA editing [26]. A role of hydrophobicity together with the size of the grooves has been shown for the proteins that interact with the quadro loops of stem–loop structures [27]. It has been shown that interaction mechanisms may be different and include specific recognition of the bases in a hydrophobic pocket, adaptive binding with GNRA motif in the large groove, and specific binding in the minor groove depending on the geometric dimensions. For the particles that recognize the signal it has been demonstrated that RNA–protein interactions occur through the specific binding with the extended large groove and the quadro loop through properly arranged water molecules without direct contact of the protein with nucleobases [28]. The roles of the small and large grooves in the interaction of the stem–loop structure with proteins have been shown for the ribosomal complex [29]. The role of the degree of the spiral twist (Twist) has also been shown for RNA–protein interactions in bovine immunodeficiency virus [30]. Hydration of the binding sites in RNA–protein interactions has been investigated for 89 RNA–protein complexes from the PDB database. It has been shown that the large groove is more hydrated than the small one in RNA–protein interactions, while the opposite dependence has been observed for the protein binding sites in DNA [31].

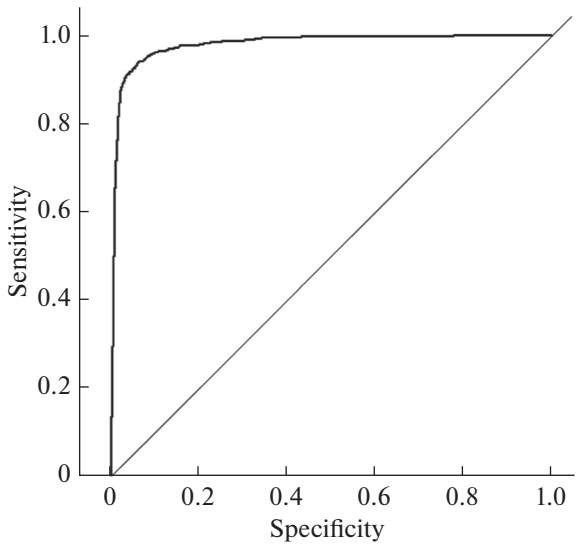

Fig. 2. The curve of the errors of the logistic regression model.

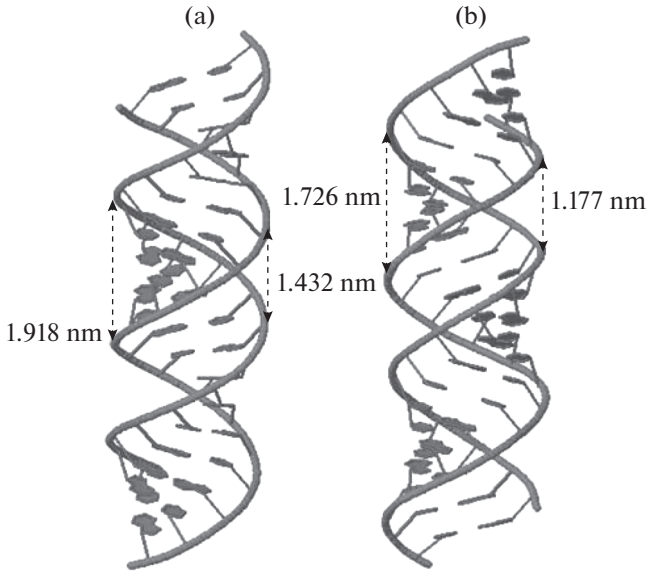

Fig. 3. (a) The model of the stem structure of a random region of the human genome; (b) the model of the stem structure located at the 3'-end of the RNA transcript of the L1 transposon in the human genome.

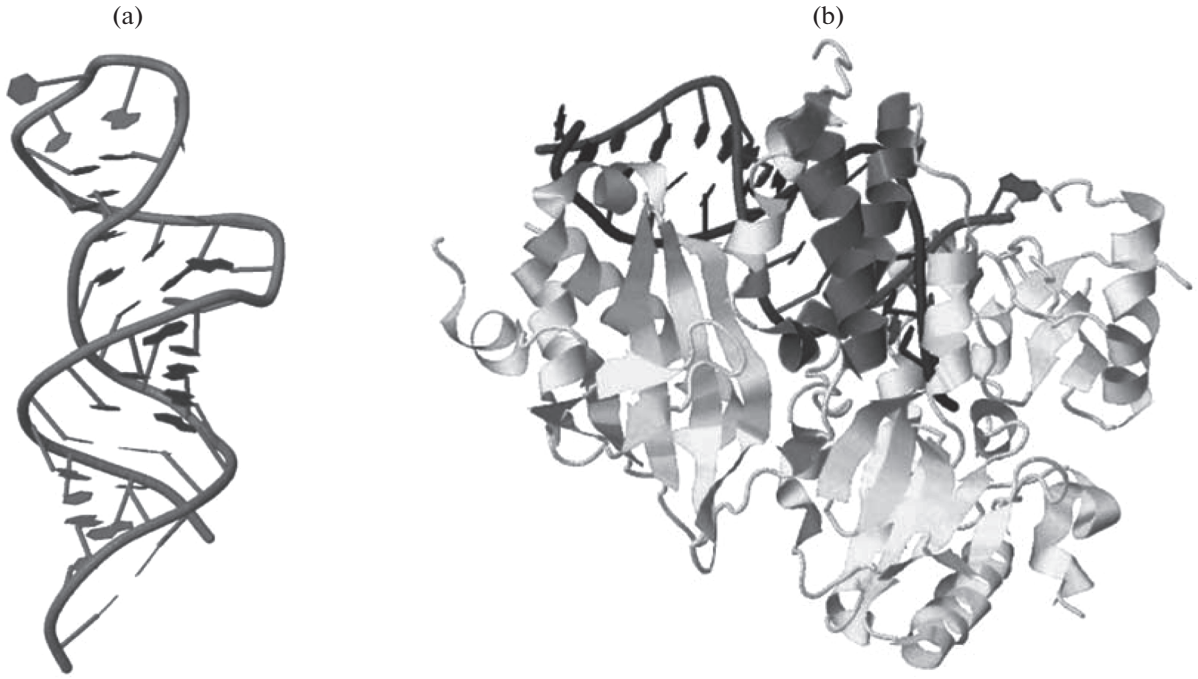

Fig. 4. (a) The hairpin structure of the LINE transposon that is recognizable by reverse transcriptase in the genome of the fish *Danio rerio*; (b) the complex of double helix RNA and retrovirus reverse transcriptase.

The model that was constructed in this work using machine learning showed that transposon stem–loop structures can be recognized by physical and geometric characteristics with an accuracy of approximately 95%; they differ from the stem–loop structures of other genome regions. The trained model also allows one to find similar structures in any given RNA sequence. It would be interesting to experimentally test the question of whether reverse transcriptase can recognize any sequence that has a secondary 3'-end structure with the properties identified in this work. The property we revealed can be of great importance in bioengineering for inserting a given sequence into the genome.

The detection of characteristics that significantly differ between groups of structures using computer methods is an important issue in the study of protein binding. If the characteristics that determine the interaction with protein are known, it may be possible to regulate this process. In this work we demonstrated the possibility of applying machine-learning methods to address this issue. The largest contribution to the difference between the transposon structures with the structures taken from the random genome regions and those generated randomly is made by four characteristics: the hydrophilicity, the enthalpy, and the Rise and Twist geometric parameters. The hydrophilicity and the Rise parameter have lower values for the transposon structures. The Twist parameter, in contrast, has the maximum value. It is assumed that the preservation of the structural properties of transposons in the group of active transposons is not accidental and that these properties and not sequences are under positive natural selection.

CONCLUSIONS

The physico-chemical and geometric characteristics of the stem–loop structures at the 3'-end of human L1 transposons influenced by natural selection have been identified. These characteristics include thermodynamic parameters such as the Gibbs free energy, enthalpy, entropy, and hydrophilicity, along with six geometric parameters of RNA structure: Shift, Slide, Rise, Tilt, Roll, and Twist. It is possible to identify stem–loop structure in any given RNA sequence with properties similar to the specified physico-chemical characteristics of the stem–loop structures at the 3'-end of the active L1-transposon using machine-learning methods. The identification of the key secondary structure characteristics of RNA binding with reverse transcriptase may have important practical applications in bioengineering for inserting a set of sequences into the genome.

REFERENCES

1. C. R. Huang, K. H. Burns, and J. D. Boeke, Annu. Rev. Genet. **46**, 651 (2012).
2. E. S. Lander, et al., Nature **409** (6822), 860 (2001).
3. D. C. Hancks and H. H. Kazazian, Jr., Curr. Opin. Genet. Dev. **22** (3), 191 (2012).

4. C. R. Beck, et al., Annu. Rev. Genomics Hum. Genet. **12**, 187 (2011).
5. H. H. Kazazian, Jr., Science **303** (5664), 1626 (2004).
6. S. R. Richardson, et al., Microbiol. Spectr. **3** (2), MDNA3-0061-2014 (2015).
7. Y. Hayashi, et al., Nucleic Acids Res. **42** (16), 10605 (2014).
8. M. Kajikawa and N. Okada, Cell **111** (3), 433 (2002).
9. Osanai, M., et al., Mol. Cell Biol. **24** (18), 7902 (2004).
10. D. Grechishnikova and M. Poptsova, BMC Genomics **17** (1), 992 (2016).
11. N. M. Luscombe, et al., Genome Biol. **1** (1), REVIEWS001 (2000).
12. P. Barraud and F. H. Allain, Curr. Top. Microbiol. Immunol. **353**, 35 (2012).
13. W. Chen, et al., Sci. Rep. **6**, 35123 (2016).
14. W. Chen, et al., Nucleic Acids Res. **41** (6), e68 (2013).
15. W. Chen, et al., Biomed. Res. Int. **2014**, 623149 (2014).
16. B. Liu, F. Yang, and K. C. Chou, Mol. Ther. Nucleic Acids **7**, 267 (2017).
17. M. Friedel, S. Nikolaiewa, J. Sühnel, and T. Wilhelm, Nucleic Acids Res. **37** (Database issue), D37 (2009).
18. X. J. Lu and W. K. Olson, Nucleic Acids Res. **31** (17), 5108 (2003).
19. C. O. Pabo and R. T. Sauer, Annu. Rev. Biochem. **53**, 293 (1984).
20. P. C. van der Vliet and C. P. Verrijzer, Bioessays **15** (1), 25 (1993).
21. R. E. Dickerson, Nucleic Acids Res. **26** (8), 1906 (1998).
22. V. A. Tverdislov, Biophysics (Moscow) **58** (1), 128 (2013).
23. G. Masliah, P. Barraud, and F. H. Allain, Cell. Mol. Life Sci. **70** (11), 1875 (2013).
24. R. Stefl, L. Skrisovska, and F. H. Allain, EMBO Rep. **6** (1), 33 (2005).
25. J. R. Williamson, Nat. Struct. Biol. **7** (10), 834 (2000).
26. J. M. Thomas and P. A. Beal, BioEssays **39** (4), 1600187 (2017).
27. R. Thapar, A. P. Denmon, and E. P. Nikonowicz, Wiley Interdiscip. Rev. RNA **5** (1), 49 (2014).
28. K. Wild, I. Sinning, and S. Cusack, Science **294** (5542), 598 (2001).
29. G. L. Conn, Science **284** (5417), 1171 (1999).
30. D. Moras and A. Poterszman, Curr. Biol. **6** (5), 530 (1996).
31. A. Barik and R. P. Bahadur, Nucleic Acids Res. **42** (15), 10148 (2014).

Translated by E. Puchkov

SPELL: 1. OK