Computer simulation of tRNA evolution

2009 J. Phys. A: Math. Theor. 42 345101

(http://iopscience.iop.org/1751-8121/42/34/345101)

IOP PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND THEORETICAL

J. Phys. A: Math. Theor. **42** (2009) 345101 (9pp) doi:10.1088/1751-8113/42/34/345101

Computer simulation of tRNA evolution

Fangping Wei[1,2], **Sheng Li**[2] **and H R Ma**[2]

[1] Institute of Theoretical Physics, GuangXi University, GuangXi 530004, People's Republic of China
[2] Institute of Theoretical Physics, Shanghai Jiao Tong University, Shanghai 200240, People's Republic of China

Received 3 June 2009, in final form 1 July 2009
Published 31 July 2009
Online at stacks.iop.org/JPhysA/42/345101

Abstract

Point mutation and complementary theory are two controversial mechanisms of tRNA evolution. To identify the predominant or most suitable mechanism to explain modern tRNA evolution, this paper presents a rough evolution model for modern tRNA sequences using computer simulation. The tRNA networks of the sequences produced by the model were compared with the real tRNA network. It was found that when the 'ancestor sequences' have a probability greater than 50% of mutating by the point mutation mechanism and a probability less than 50% of mutating by the complementary anticodon method at the same time, the tRNAs produced by the model should have a network behavior similar to that of the real tRNAs network.

PACS number: 87.10.Vg

Introduction

Two controversial mechanisms have been proposed to account for tRNA evolution: point mutation and the complementary method. Both these mechanisms are supported by many theories and models [1–7]. Point mutation is based on the assumption that a tRNA gene can be recruited from one isoaccepting group to another by a point mutation such as nucleotide replacement, insertion, deletion and duplication, which concurrently changes the tRNA amino acid identity and the messenger RNA coupling capacity. In 1998, Saks *et al* found that an Escherichia coli strain, in which the essential $tRNA^{Thr}_{UGU}$ gene was inactivated, was rendered viable when a $tRNA^{Arg}$ with a point mutation that changed its anticodon from UCU to UGU (threonine) was expressed [2]. The complementary method suggests an evolutionary scheme for the oldest tRNA sequences using a hypercycle theory [6] which states that the ancestors of modern tRNAs appear to have emerged by the shortest possible way, both complementary strands of a short symmetrical double helix serving as pre-tRNAs with complementary anticodons [3]. This theory proposes that four initial pairs of pre-tRNA with complementary anticodons are capable of generating a total of 64 anticodons. These two

evolutionary mechanisms for modern tRNAs are all reasonable from the theoretical standpoint, but which is the most suitable mechanism for describing tRNA evolution? This paper compares the degree of distribution and the clustering coefficient of networks constructed by the tRNA sequences of the single anticodon group, single isoaccepting group, and the whole tRNA group of parallel and antiparallel networks. The result of this comparison seems to be consistent with the idea that modern tRNA sequences evolved primarily by the mechanism of the complementary method, but that point mutation is an important, indispensable, and complementary mechanism during the evolutionary process [8]. This paper undertakes further investigation of this theory. A simple model for tRNA evolution using computer simulation is presented, and the tRNA network of sequences produced from the model is compared with the real tRNAs network. It was found that when the 'ancestor sequences' have a probability greater than 50% of mutating with the point mutation mechanism and a probability less than 50% of mutating with complementary anticodon method at the same time, the tRNAs produced by the model should have network behavior similar to that of the real tRNA network.

1. Material and methods

1.1. The model for the tRNA evolution

The model was designed as follows:

(1) Assume that there are a_0 tRNA sequences which serve as the evolutionary precursors or evolutionary seeds. Set $b_0 = a_0$.
(2) Assume that each tRNA sequence accumulates mutation and becomes two sequences. So, the number of the tRNAs should be evolved as follows:

$$\begin{aligned} &b_0 = 2^0 a_0 - c_0 (c_0 = 0) \\ &b_1 = 2^1 a_0 - c_1 \\ &\dots \\ &b_n = 2^n a_0 - c_n \\ &(n \text{ is the mutation time steps}, n = 0, 1, 2, \dots, N) \end{aligned} \tag{1}$$

In equation (1), c_n is the number of tRNA sequences which evolve from b_{n-1} but are then eliminated by natural selection during the nth mutation time step. b_n is thus the surviving number of tRNA sequences.

Let us consider the technical details of how to put the model into practice. The first question is how to select the evolutionary precursor. The approach used here is based on the following reasoning. (1) the precursor sequence must come from the tRNA family that accepts the simplest amino acid. It is said that simple molecular building blocks, such as proteins and cells, can come together to form complex organisms. Consequently, the amino acid should have the simplest possible molecular structure [9], and the sequence and structure of primordial RNA may have originally carried amino acids [10]. Different amino acids have different side chains, which leads to their diversity. Among the side chains of the 20 amino acids, glycine (Gly) has the simplest side-chain structure; it contains only a hydrogen atom (H). Thus, the tRNA sequence that accepts the message of Gly is to be considered as the evolutionary seed. (2) The ancestral sequence must be the most conservative or 'fossil' sequence among the modern tRNA sequences [11, 12]. (3) Its secondary structure must yield the prototypical tRNA structure. For these purposes, 38 tRNA sequences of eight species, including three kingdoms: eukaryotes, prokaryotes and viruses, were retrieved from the Sprinz database [5]. These tRNAs come from the same tRNA isoacceptors with the amino-acid identity of

Gly, representing four anticodon subsets: GGA, GGC, GGG and GGT. Each tRNA sequence consists of 99 bases. To find the optimal candidate for the precursor, these 38 tRNAs were aligned; two tRNAs with anticodon GGA were found to be the most conservative, and one of them was selected as the precursor sequence.

Then a probability ξ is defined to determine which type of evolutionary mechanism each seed sequence should select at the beginning of mutation. ξ is a stochastic decimal variable ranging between 0 and 1. When $\xi = 0$, this means that evolution occurs using only the point mutation mechanism. When $\xi = 1$, this means that evolutionary events occur using only the complementary-method mechanism. When $1 > \xi > 0$, this means that evolution takes place using a mixture of the two mechanisms, with ξ of the possible seed sequences evolving using the complementary-method mechanism and $1 - \xi$ of the sequences evolving using the point-mutation mechanism.

When using the point mutation mechanism, four possible point mutations occur randomly on the precursor sequence. One is nucleotide substitution, meaning that some nucleotides are replaced by other bases, such as A substituting for T or G or C in the homologous site. The second is nucleotide deletion, meaning that some nucleotides of the sequence are deleted. The third is nucleotide insertion, meaning that some positions of the sequence are inserted into nucleotides. The final one is nucleotide duplication, meaning that some nucleotides are repeated in the same site, for example replacement of A by AA. These four events happen in the sequence under the condition that the total number of bases in the tRNA sequences remains unchanged. When using the complementary mechanism,the seed sequence acts as a template to duplicate the other tRNA strand with complementary bases, such as $A \longrightarrow U$ or $C \longrightarrow G$, or $U \longrightarrow A, G \longrightarrow C$ or $G \longrightarrow U$. Five sites, 0, 73, 74, 75 and 76, remain unchanged throughout the duplication.

Not all tRNA sequences that accumulate mutations from a_0 can survive, only the sequence and structure of those tRNAs agree with the requirements of natural selection can be conserved. The estimation criterion is based on three conditions. (1) the structure of the new sequence content contains the cloverleaf secondary structure, and the number of Watson–Crick base pairs, ϕ, lies in the range $15 \leqslant \phi \leqslant 22$ [12, 13]. (2) If the mutation occurs only in the anticodon position, the new sequence should be assigned a probability of survival after competing with the evolutionary seed sequence. (3) Each new tRNA sequence which produced from the n th mutation time step should not be the repetitive sequence of the tRNA which have been survival before. Even if the new sequence fulfills these three conditions, it is then subjected to an attenuation survival probability, p_{survival}:

$$p_{\text{survival}} = \beta \exp(-\alpha n) \begin{cases} \beta = 1, & \text{when} \quad 18 \leqslant \phi \leqslant 22 \\ 0 < \beta < 1, & \text{when} \quad 15 \leqslant \phi < 18, \quad \text{and} \quad 22 < \phi \leqslant 25 \end{cases} \tag{2}$$

where α and β are constant variables, and n is the mutation time step index. p_{survival} is a probability of evolution that is related to the struggle for survival between the new sequences and the old seed sequences. Viewed from an evolutionary perspective, old species have a greater probability of surviving than new species, so most newcomers eventually disappear, with only a few surviving. The total number of a given species should eventually trend to equilibrium. The equation for p_{survival} is deduced from this view.

Five parameters can therefore be used to calibrate the model: p_{key}, m, α, β and ξ. p_{key} is the probability that a new sequence will survive when the seed sequence and the new sequence are the same except for the anticodon sets. From the definition, one can easily conclude that the value of p_{key} should influence the number of anticodons in the tRNA. If the value of p_{key} is too small, few new tRNA sequences with different anticodons can survive, and anticodon multiformity in the tRNA sequences will be destroyed. The second parameter, m, represents

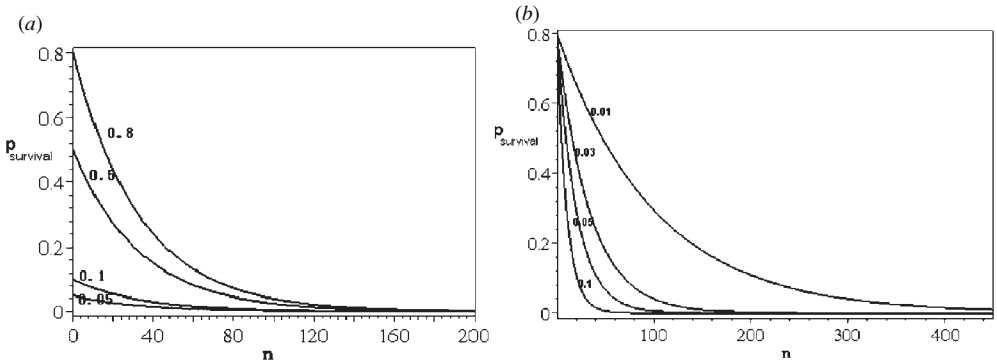

Figure 1. The distribution of p_{survival}. (*a*) Fixed $\alpha = 0.03$, β changed within the value of 0.05, 0.1, 0.5 and 0.8; (*b*) fixed $\beta = 0.5$, α changed within the value of 0.01, 0.03, 0.05 and 0.1.

how many sites on the seed sequence experience mutation in each time step. Previous work has estimated that the entire avian mitochondrial genome undergoes substitutions at a rate of 0.02 mutations per site per million years (s/s/myt) [14] or a rate of 0.4 to 1.4 s/s/myt for mitochondrial HVRI [15]. Therefore, m should have a value of fewer than 10 sites per time step, because rough calculations reveal that if m becomes greater than 10, the new sequences will suddenly become silent, the secondary structure of the new tRNA sequences will be unable to achieve the cloverleaf structure of the consensus tRNAs, and the tRNA sequences will eventually become completely random sequences. The third and fourth parameters, α and β, are both attenuating constant variables ranging between 0 and 1. α, β and the mutation time step n work in common to influence p_{survival} as shown in equation (3). Figure 1 shows that when $\alpha n \longrightarrow$ a large value, then $p_{\text{survival}} \longrightarrow 0$. When $p_{\text{survival}} \longrightarrow 0$, the model should converge to a fixed constant or an equilibrium state. Consequently, the model is influenced by three parameters: p_{key}, p_{survival} and ξ. p_{key} determines the degree of multiformity of the tRNA sequences, p_{survival} determines the scale and the equilibrium point of the model, and ξ determines the behavior of the end products of the model. All the above analysis is purely theoretical; the final result of the model is also influenced by natural selection, which is simulated by a group of well-behaved random decimal variables ranging between 0 and 1.

1.2. The parameters impact on the model

Figure 2 shows how the model is influenced by p_{key}, m, α, β and ξ. A common characteristic of the curve b_n in figure 2 is, when p_{survival} tends to a very small value, the state of the model reaches equilibrium. For a fixed value of α, mutation time step, n, divided b_n into three zones. In the first, b_n increases quickly with low value of n; the second shows a slight increasing trend among values close to $\alpha n \sim 10$. The third zone of b_n represents a steady state, in which, the scale of b_n remains unchanged.

1.3. Result

The model may not be able to generate tRNA sequences identical to real tRNA sequences in nature, but the generated sequences have some characteristics similar to those of real sequences, such as the primary and secondary structures of the tRNA, the equilibrium of the final result of the evolution, and the indeterminate outcome of the mix of tRNA sequences.

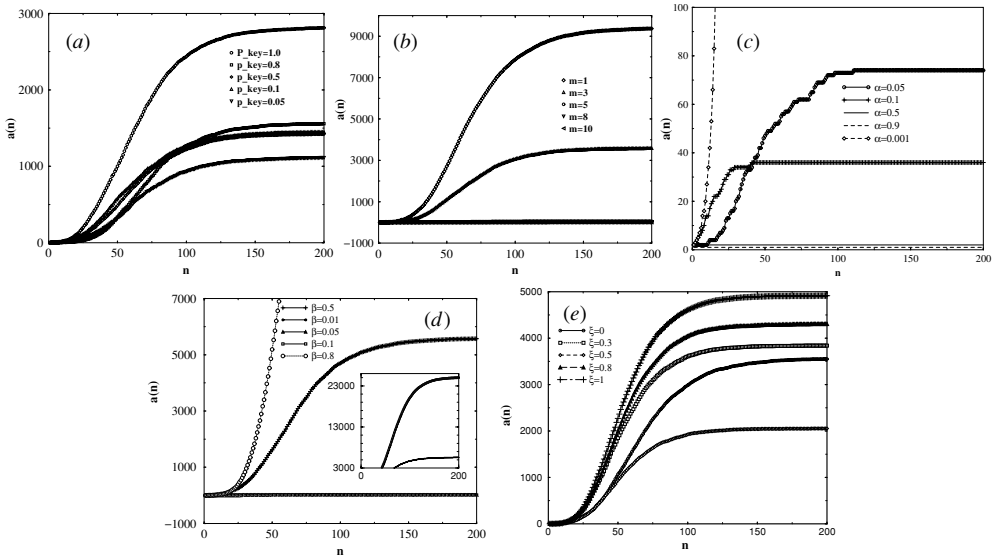

Figure 2. p_{key}, m, α and β influence on the model. (a) Fixed $m = 2, \alpha = 0.04, \beta = 0.5$, $\xi = 0$, p_{key} changed within the value of 0.01, 0.05, 0.1, 0.5 and 1.0; (b) fixed $p_{\text{key}} = 0.5$, $\alpha = 0.04, \beta = 0.5, \xi = 0$, m changed within the value of 1, 3, 5, 8 and 10; (c) fixed $p_{\text{key}} = 0.5, m = 2, \beta = 0.5, \xi = 0$, α changed within the value of 0.01, 0.05, 0.1, 0.5 and 0.9 (The curve of $\alpha = 0.001$ only draw $n \leqslant 40$ steps, for $n > 50$, it beyond the charge of personal computer.); (d) fixed $p_{\text{key}} = 0.5, m = 2, \alpha = 0.04, \xi = 0$, β changed within the value of 0.01, 0.05, 0.1, 0.5 and 0.8, the insets in (d) show a sequel of curve $\beta = 0.8$; (e) fixed $p_{\text{key}} = 0.5, m = 2, \alpha = 0.05, \beta = 0.5$, ξ changed within the value of 0, 0.3, 0.5, 0.8 and 1. b_n record the number of the tRNAs survive in n mutation time steps.

Many biological systems can be described as networks with complex topologies. A network is made up of a set of nodes(or call vertices) and connections between them(call edges or links or connections). The feature and nature of the network are indicated mainly by two parameters, degree k and clustering coefficient c of nodes. The degree k of a node is the number of the other nodes to which it connects. Basically, the networks can be classified into two types in terms of its degree distributions $p(k)$ of nodes: exponential networks and scale-free networks. The former type has a prominent character that almost every node has the same number of edges, $k \sim \langle k \rangle$. This distribution leads to a Poisson or exponential distribution. The latter type of network has a feature that few nodes have many links, but many nodes have few links. Its degree distribution appears a power-law distribution, $p(k) \sim k^{-\gamma}$. It is also called inhomogeneous networks, or scale-free networks. If a node connects with i other nodes and there are j edges connected within these i nodes, the clustering coefficient c of the original node is defined as

$$c = \frac{2j}{i(i-1)}, \tag{3}$$

where $i(i-1)/2$ is the total number of possible connections among i nodes. The clustering coefficient reflects relationships of the neighbors of a node, and quantifies the inherent tendency of the network to clustering. To construct the tRNA similarity network, each tRNA sequence is considered as a node. If the alignment score s (alignment score is defined as the number of bases in the home sites of the two tRNA sequences which are the same in parallel comparison, such as A–A, G–G, T–T and C–C. Similarity degree is defined as $s = \frac{\text{matching scores}}{L} \times 100$. L is the total number of the nucleotides of the sequence) is larger than a given similarity

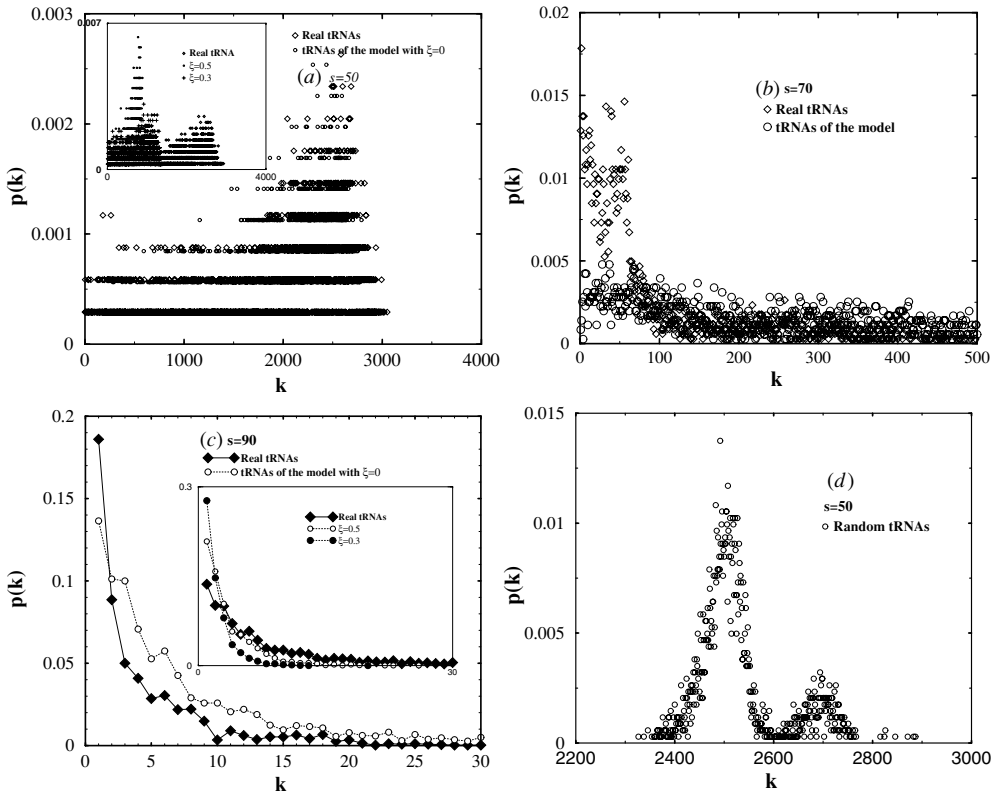

Figure 3. The degree distribution of the network of the model tRNAs, real tRNAs and its corresponding random tRNAs. Main panel: the degree distribution of the network of the model tRNAs with $\xi = 0$ and real tRNAs. Inset: the degree distribution of the network of the model tRNAs with $\xi = 0.3$ and $\xi = 0.5$ and real tRNAs.

degree s_0, the corresponding nodes are linked. So, an undirected complex tRNA similar network is constructed [16, 8]. It was found that, when the degree of similarity $s > 70$, the distribution of notes $p(k)$ of the network appears to follow a power-law distribution, and its corresponding random tRNA network appears a Gauss distribution for $s < 70$. When $s > 70$, the random tRNA network $p(k) \longrightarrow 0$ [16]. Figure 3 shows the distribution of the degree of similarity of the three networks, one constructed using the tRNA sequences produced by the evolutionary model described in this paper, and the others constructed using the real tRNAs and their corresponding random tRNAs. As can be observed in figure 3(a), when $s = 50$, the distribution curves of the tRNA sequences produced by the model with $\xi = 0, \xi = 0.3$ and $\xi = 0.5$ behave similarly to the curve of the real tRNA sequences, except that the peaks of the curves depart somewhat from that of the curve for the real tRNAs. All the curves appear to have a poorly defined shape. With increasing ξ, it is uncommon for nodes of the network derived from the model to have a degree of less than 1 compared with those derived from the real tRNAs. When $s = 70$, the two curves begin to appear to follow a power-law distribution, except that in the region $k \sim 10$ to 30, the two curves deviate slightly. Compared with the network derived from random tRNA sequences (see figure 3(d)), the network derived from the model developed here shows clear differences in behavior. As the degree of similarity changes from 50 to 60, the behavior of the distribution of the random tRNA network always appears

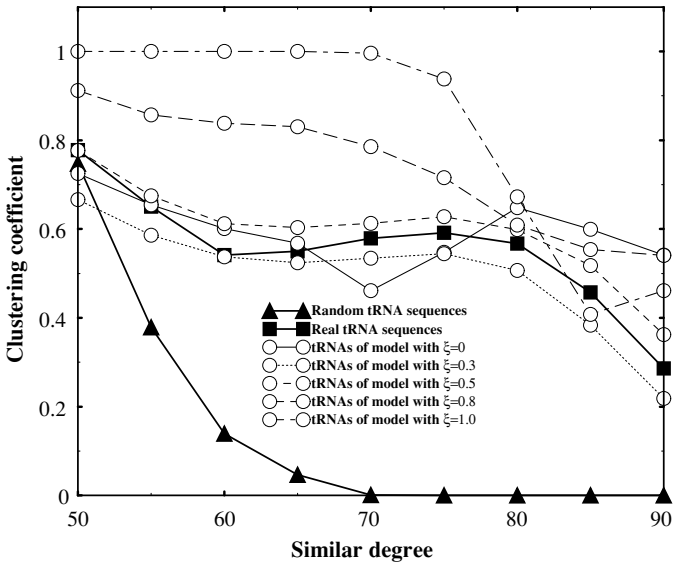

Figure 4. The distribution of the clustering coefficient of the three networks according to their similar degree.

Table 1. The average clustering coefficient of the tRNAs network of our model with $\xi = 0, \xi = 0.3, \xi = 0.5, \xi = 0.8$ and $\xi = 1.0$, and the real tRNAs network and its corresponding random tRNAs network. The numerical value in the bracket is the value of ξ of the model.

s	$c_{\text{model}}(0)$	$c_{\text{model}}(0.3)$	$c_{\text{model}}(0.5)$	$c_{\text{model}}(0.8)$	$c_{\text{model}}(1.0)$	c_{real}	c_{random}
50	0.725 485	0.666 534	0.777 108	0.911 686	1.000 000	0.777 367	0.747 479
60	0.601 328	0.537 331	0.612 162	0.838 526	1.000 000	0.541 708	0.139 572
70	0.461 4946	0.534 485	0.613 036	0.785 719	0.996 391	0.578 806	0.000 682
80	0.648 3536	0.506 975	0.598 649	0.608 761	0.672 637	0.567 380	0.000 000
90	0.540 9976	0.218 362	0.362 006	0.540 537	0.461 549	0.286 254	0.000 000

to follow a Gaussian distribution. When $s = 70$, most vertices of the random tRNA network have lost their connection. When $s = 90$, figure 3(c) shows that the three curves of the model with $\xi = 0, \xi = 0.3$ and $\xi = 0.5$ behave identically to the distribution of the real tRNAs, appearing to follow a power-law distribution. In summary, the tRNA sequences evolved from the model developed here provide a discrimination of random tRNAs; their behavior is more like that of the real tRNAs than of the random ones, which means that the tRNA sequences generated by this model have similar overall network characteristics to those of real tRNAs.

This observation can also be made about the other curves with $\xi \leqslant 0.5$. When $\xi > 0.5$, the behavior of the model begins to deviate from the behavior of the real tRNAs. Especially when $\xi = 1$, the overall network of the model shows little resemblance to the behavior of the real tRNAs, but it is still clearly different from the network of the random tRNAs.

Another important parameter, the average clustering coefficient, c_{model}, of the network should also be considered. As can be observed in table 1 and figure 4, c_{model} is larger than the average clustering coefficient c_{random} of the random tRNAs, except for $c_{\text{model}}(0.3)$ in the region $s = 50$. The distribution of c_{model}, unlike the distribution of c_{random}, drops from a comparatively

large value directly to zero. It first decreases when $s < 70$, then increases slowly after $s > 70$, but when $s > 80$, it drops down again. These behaviors are similar to those of real tRNA sequences. Among the curves of c_{model}, two curves with $\xi = 0.3$ and $\xi = 0.5$ showed behavior most similar to that of real tRNAs; there are only minor discrepancies between the real and model curves, such as the fact that in the region $50 \leqslant s \leqslant 60$, the distribution of $c_{\text{model}}(0.3)$ and $c_{\text{model}}(0.5)$ is smoother than that of c_{real}. When $\xi > 0.5$, c_{model} showed more marked deviations from c_{real}, and the peak of c_{model} shifted to the left at $s = 70$. It was also found that the average clustering coefficient of all 20 group tRNAs with $\xi \leqslant 0.5$ behaved similarly to the 20-group clustering coefficient of the real tRNAs in [8], which means that the network generated by the present model with $\xi \leqslant 0.5$ has local characteristics similar to those of the real tRNAs network.

2. Conclusion

This paper has presented a rough evolutionary model for modern tRNA sequences, based on point mutation theory and the complementary method. The evolutionary mechanisms of the model rely on the assumption that there exists an oldest tRNA sequence to be the seed of the evolutionary event. The ancestral sequence accumulates mutations, such as nucleotide replacement, nucleotide insertion, nucleotide deletion and nucleotide duplication, becoming the newcomer in the tRNA gene families. The mutated ancestral sequence then undergoes selection by the 'survival of the fittest' rule of natural selection, which is emulated by an attenuation survival probability, $\beta \exp(-\alpha n)$. Finally, the survivors join the tRNA family and become the seed of the next generation. Such an evolutionary mechanism leads to a result which behaves similarly to natural selection with $0 \leqslant \xi \leqslant 0.5$. At the same time, the model is in an equilibrium state when $\alpha n \longrightarrow$ a large value, it finally converges to a constant value and then remains in a steady state. Networks constructed using tRNA sequences randomly generated by the model developed here under the condition $0 \leqslant \xi \leqslant 0.5$, and with a number of sequences comparable to the real tRNAs, behave similarly to a real tRNA network. The node distribution of the networks shifts from an indefinite shape to that of a power-law distribution when $s > 70$, which provides clear discrimination from the random tRNA network. The average clustering coefficient of the tRNAs produced from the present model also behaves very differently from the coefficient of a random tRNA network. A prominent characteristic of the average clustering coefficients of the real tRNA sequences is that the distribution curve exhibits a peak before dropping down to zero; the distribution curve of the average clustering coefficients from the model also shows a peak before decreases to zero. Not only do the global characteristics of the network developed from the model ($0 \leqslant \xi \leqslant 0.5$) behave similarly with the real tRNA network, but also the local characteristics of the two networks are alike, as manifested by the clustering coefficient of the 20 group tRNA sequences of the two networks classified by their amino-acid identities. Two of the networks show a clear difference from the random tRNA network. On the other hand, when $0.5 < \xi \leqslant 1$, the network derived from the model deviated markedly from the real tRNAs network, which means that the evolutionary mechanism of modern tRNA must not rely on the complementary method alone. Is it possible that the point-mutation mechanism alone can account for modern tRNA evolution? This mechanism seems to be too slow and uneconomical to produce the observed multiformity of modern tRNA sequences. It would therefore appear that a rational evolutionary mechanism for the tRNAs as deduced from the model would be a mixture of the two mechanisms. To determine which of the two is predominant during the evolutionary event, further research on the model is needed.

Acknowledgments

This paper was supported by the GuangXi National Science Foundation under Grant No.0728003.

References

[1] Margaret E S, Jeffrey R S and John A 1998 Evolution of a transfer RNA gene through a point mutation in the anticodon *Science* **279** 1665–70
[2] Saks M and Sampson J R 1998 Evolution of tRNA recognition system and tRNA gene sequence *J. Mol. Evol.* **40** 509–18
[3] Sergey R, Susumu O and Andrey R 1993 Transfer RNAs with complementary anticodons: could they reflect early evolution of a discriminative genetic code adaptor? *Proc. Natl Acad. Sci. USA* **90** 4723–7
[4] Nancy M and Alan M W 1995 Tempo and mode in evolution: genetics and paleontology 50 years after Simpson *Nucl. Acids Res.* 25–41
[5] Sprinz M, Horn C, Brown M, Loudovith A and Steinberg S 1998 Compilation of tRNA sequences and sequences of tRNA genes *Nucleic Acids Res.* **26** 148–53
[6] Schuster P 1993 RNA-based evolutionary optimization *Orig. Life* **23** 373–91
[7] Schultz D W and Yarus M 1994 tRNA Structure and ribosomal function: I. tRNA nucleotide 27–43 mutationsenhance first-position wobble *J. Mol. Biol.* **235** 1381–94
[8] Wei F P, Meng M, Li S and Ma H R 2006 Comparing two evolutionary mechanisms of modern tRNAs *Mol. Phylogenet. Evol.* **38** 1–11
[9] Ulrich B and Oro J 1993 Three stages in the evolution of the genetic code *Biosystems* **29** 133–41
[10] James F C 1973 *Evolution: Possible or Impossible* (Northridge, CA: Probability Research in Molecular Biology)
[11] Hopfield J J 1978 Origin of the genetic code: a testable hypothesis based on tRNA structure, sequence, and kinetic proofreading *Proc. Natl Acad. Sci. USA* **75** 4334–8
[12] Yasubumi S, Michael B, Richard H, Mian I S, lander S K, Rebecca C U and Haussler D 1994 *Recent Methods for RNA Modeling Using Stochastic Context-Free Grammars* (Berlin: CPM Springer) 289–306
[13] Stefan W, Walterna F, Ivo L H and Peter S 1999 Complete suboptimal folding of RNA and the stability of secondary structures *Biopolymers* **49** 145–65
[14] Shields G F and Wilson A C 1987 Calibration of mitochondrial DNA evolution in geese *J. Mol. Evol.* **24** 212–7
[15] Lambert D M, Ritchie P A, Millar C D, Holland B, Drummond A J and Baroni C 2002 Rate of evolution in ancient DNA from Adlie Penguins *Science* **295** 2270–3
[16] Wei F P, Li Y and Ma H R 2008 The network of tRNA gene sequences *J. Shanghai Jiaotong University (Science)* **13** 611–6