SCIENCE CHINA
Life Sciences

• INSIGHT •

April 2017 Vol.60 No.4: 432–435
doi: 10.1007/s11427-016-9011-7

A natural communication system on genome evolution

Qi Wu[1], Yadi Wang[2], Yun Ding[1,3], Shuai Ma[1,4], Zongmin Wu[2*] & Fuwen Wei[1*]

[1]*Key Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China;*
[2]*Shanghai Key Laboratory for Contemporary Applied Mathematics, School of Mathematical Sciences, Fudan University, Shanghai 200433, China;*
[3]*School of Life Sciences, Nanchang University, Nanchang 330031, China;*
[4]*College of Life Sciences, University of the Chinese Academy of Sciences, Beijing 100049, China*

Received December 6, 2016; accepted December 22, 2016; published online March 13, 2017

Citation: Wu, Q., Wang, Y., Ding, Y., Ma, S., Wu, Z., and Wei, F. (2017). A natural communication system on genome evolution. Sci China Life Sci 60, 432−435. doi: 10.1007/s11427-016-9011-7

The central dogma of molecular biology describes the flow of genetic information from DNA via RNA to protein and duplication from ancestral to descendent DNA (Crick, 1958). However, the genetic information could not be quantified and mathematically modeled. So it differs from the "information" formulated by Shannon and used in information and coding theories (Shannon, 1949). Although physicists have suggested that life absorbs negative entropy (or information) from the environment (Schrodinger et al., 1944), no physical model has described the transmission of information during evolution. Similarly, while biological researchers focus on variations in biological characteristics, they pay less attention to the relationship between the characteristics and time. Thus, they consider evolution as a means of pure classical mechanics. In the present study, we attempt to construct a framework of genome evolution based on the potential relationship of time and information which we can use to discuss the possible links between information changes of characteristic variation and the time when these variations occur.

Current artificial communication systems (ACSs) involve two different spatial points and the fundamental process of either exactly or approximately reproducing a message from the first point at the second point 0. Evolution can be analogously described as involving two temporal points and the fundamental process of either exactly or approximately reproducing genetic information from the former point at the latter point. However, one difficulty with this analogy is supposing the existence of a channel between ancestral and descendent species permitting information to be encoded and transferred. To address this problem, we introduce time into the concept such that the channel for information transmission during evolution is no longer a material, space-occupying element, but a type of transactional, time-occupying one. The channel is transactional because the duplication of genetic information (together with variation-producing novel information) during speciation is the only element connecting ancestral and descendent species. The channel is also time-occupying because its performance can be measured by the time taken to pass through the channel, which is referred to as channel time. Therefore, the model of a natural communication system (NCS) for evolution consists of a material source, a material destination, and a transactional channel connecting them and producing heritable variation. The entire process involves the copying (with variation) of genetic information of entire genomes from ancestral species resulting in descendent species containing evolutionary novelties (Figure 1).

In an ACS, symbols at the source cannot be directly loaded or transferred into the channel, but must instead be encoded

*Corresponding author (Zongmin Wu, email: zmwu@fudan.edu.cn; Fuwen Wei, weifw@ioz.ac.cn)

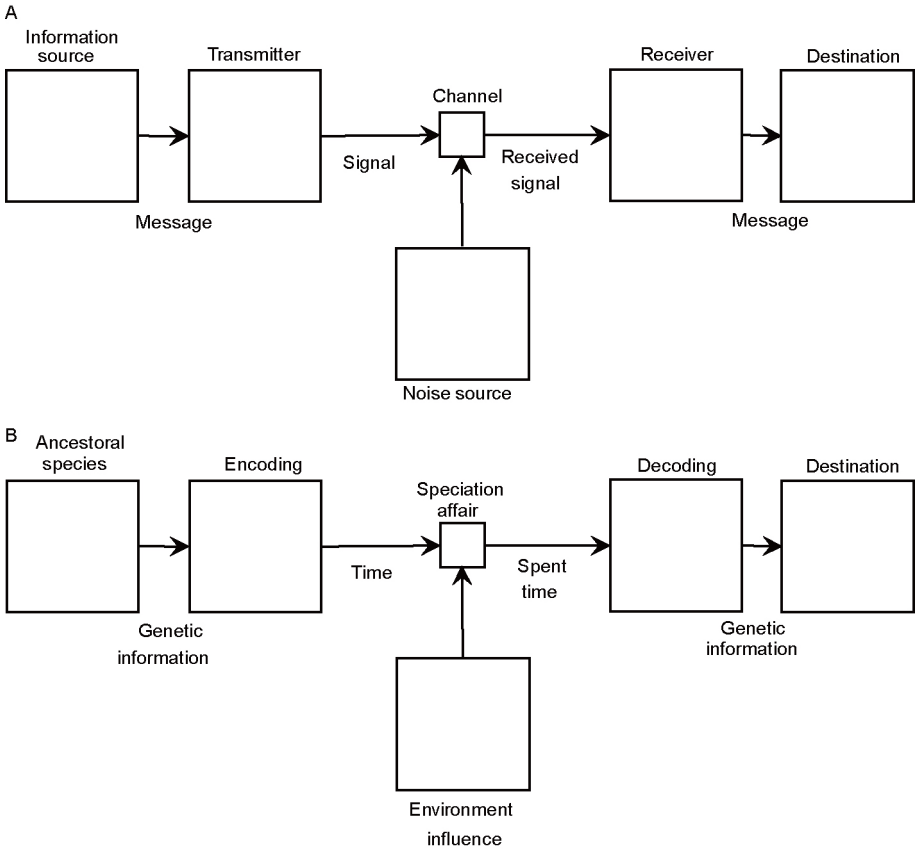

Figure 1 Comparison of an ACS and a NCS. A, An ACS modified from Shannon's classical schematic diagram (Shannon et al., 1949). B, NCS. Note that the genetic information encoded and transferred are the genetic information of the entire genome.

as channel codes that can be transferred directly to the channel, called source encoding 0. In the source coding theory, the highest efficiency is able to be achieved by source encoding 0. Similar cases are seen in an NCS. Evolution is driven by selection, as is the evolution of different information transmission processes in NCSs. Because an information transferring process with a selective advantage is more likely to be preserved, life systems had to evolve higher source encoding efficiency to attain evolutionary advantages during the evolution of NCSs. We can assume that present NCSs are bound to maintain source encoding efficiency close to the highest level permitted; We termed this as the assumption of high efficiency of NCSs.

In describing the time needed for the symbol encoded by the uniquely decodable codes to be transferred to the channel, a typical expression would be that non-noise discrete channels transfer one code character in t seconds of time. Let us name the persistence of time as a type of perennity of time and study the background frame or premise hiding behind the text. The unit of "second" comes naturally from the unit of "day", which is the perennity of time for one circle of the earth's rotation. Implicit here is that the perennity of time for a uniquely decodable code is measured by the perennity of time for one earth rotational cycle. Other measurements can be used to determine time (a case in astronomy is that one second is defined by the energy level transition period of the ground state of the cesium-133 atom), but they all use a motion in space to measure the perennity of time. Time, space, and information are all basic physical quantities. The above premise indicates that space is always used to measure time which then measures information as space-measured time. In discussing the relationship between time and information in NCS, we attempted to use the average length of a uniquely decodable code instead of space to measure the perennity of time. Thus, the length of a uniquely decodable code of n_i could be understood as the channel time of t_i for the encoded source symbol i passing through the channel from source to destination. We remark on equation (1) with:

$$p_i \geq r^{-n_i} = r^{-kt_i}, \tag{1}$$

where k is the constant coefficient parallel to the different source encoding efficiency. Solving channel time from the equation yields:

$$t_i \geq -\frac{1}{k\log r}\log p_i = \frac{1}{k\log r}I_i, \tag{2}$$

This is the relationship between channel time and source probability of a given source symbol. When the symbol is encoded with the highest efficiency, the formula is balanced.

One essential difference with ACSs is the disparity between information in the source and destination. For a given symbol i, the probability in the source of p_S could differ from the probability in the destination of p_D. Which probability determines channel time, or is there another method to describe channel time with channel probability? To solve this problem, it is necessary to reconsider what kind of message, or which kind of quantification measurement, is transmitted in NCSs. The genome sequence consists of four kinds of nucleotides: A, C, G and T. One may regard a genome sequence as the source (or destination) with four symbols, the frequencies of which describe the statistical features of the genome. However, consider the complexity of a real genome sequence; a more practical measurement should consider the joint probability of two nucleotides, i and j, with a distance, k, along the sequence. Despite differences in definitions, similar approaches have been introduced to quantify the heterogeneity of genomic DNA sequences (Karlin and Ladunga, 1994), to construct phylogenic relationship in certain taxa (Song et al., 2014; Xu and Hao, 2009), and to formulate mechanisms of genome evolution (Luo et al., 1998). In this paper, we use the measurement of mutual information of sequences. The mutual information function can be defined as:

$$I = \sum_i \sum_j x_{ij} \log \frac{x_{ij}}{x_i x_j}, x_i, x_i = (M_1)_i x_{ij} = (M_2)_{ij}, \tag{3}$$

subject to

$$\begin{aligned} &1)\ 0 < x_{ij} < 1 \\ &2)\ \sum_i \sum_j x_{ij} = 1 \\ &3)\ x_i = \sum_j x_{ij} + \sum_k x_{ki},\ \ i, j, k = 1, 2, 3, 4; \\ &\text{and } x_{ij}, x_{ki} = (M_2)_{ij}, \end{aligned} \tag{4}$$

in which i, j, and k are A, C, G, and T represented by 1, 2, 3, and 4, respectively. Details are given in Supplementary Information section 1.1. The mutual information function is a nine-dimensional curved surface in a 10-dimensional space, named the mutual information curved surface (MICS). One genome with a given distribution of 10 nucleotide pair frequencies corresponds to one point in the MICS. During evolution, the information transmission process from the source genome to the destination genome is now modeled as a point representing the source moving to a point representing the destination on the MICS. Given the assumption of high efficiency, the movement must find the shortest distance on the curved surface. The geodesic connecting the source and destination points on the curved surface could be the best choice, which is the quantified channel from the source to the destination. The arc length is a dimensionless quantity in probability space, which determines a channel time, based on equation (3). This quantity of time is the time taken for the ancestral genome to duplicate itself to produce its descendent genome including variations, which could be expressed as:

$$t = -\frac{1}{k \log r} \log L, \tag{5}$$

where k is a constant, and L is the arc. The differential form is:

$$\mathrm{d}t = \frac{1}{k \ln r} \times \frac{1}{L} \mathrm{d}L. \tag{6}$$

We define the variable quantity of genome information as the difference between information in the source and destination. Time is defined as the channel time that the information spends to pass through the transactional channel, and the genome information evolutionary rate of μ is defined as:

$$\mu = \frac{I_D - I_S}{t_{S \to D}} = k \log r \frac{I_D - I_S}{-\log L}, \tag{7}$$

where I_S is the information in the source, I_D is the information in the destination, and $t_{S \to D}$ is the time spent from point S to point D passing through the channel. The channel was the geodesic in the MICS. The differential form of the evolutionary rate is:

$$\mu = \frac{\mathrm{d}I}{\mathrm{d}t} = k \ln r L \frac{\mathrm{d}I}{\mathrm{d}L}. \tag{8}$$

Equation (8) indicates that the evolutionary rate, μ is determined by the location in channel L, the increment of information dI, and the arc length of the geodesic dL. Since the existence of L, it could be deduced that the evolutionary rate tends to increase when an ancestor has evolved into a descendent. Suppose one ancestor has two descendent genomes. The two descendants may have different time respectively (Schmidt-Nielsen, 1984) even if they have the same mutual information variation, since their geodesic lengths may be different.

We collected 94 genomes from public databases (Tables S1 and S2 in Supporting Information) covering multicellular animals, green plants, some multicellular fungi, and some unicellular eukaryotes. We calculated the evolutionary rate using these genome data and our model. Since the unavailability to infer the statistical states of the ancestral genome (the frequencies of two nucleotides in the ancestral genome), we calculated the evolutionary rate considering the virtual evolutionary processes of every species as a virtual ancestral species and the other 93 species as virtual descendent species. The results showed that nearly 20% of the processes share a one order of magnitude higher rate difference (Figures S1 and S2; Table S3 in Supporting Information), which could be explained by the evolutionary pattern of punctuated equilibrium (Eldredge et al., 1972). When focusing on one species

as an ancestor, it can be seen that the greater the phylogenic distance between two descendent species, the greater the difference between the two destination species.

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *We thank Prof. Liaofu Luo in the Inner Mongolia University, and Yi Tao in the Institute of Zoology, Chinese Academy of Sciences for comments on this research, and Xiangjiang Zhan in the Institute of Zoology, Chinese Academy of Sciences for advice on genome data. We are grateful to the Supercomputing Center of the Chinese Academy of Science (SCCAS) for the assistant and help in program coding and computation. This work was supported by the National Natural Science Foundation of China (31372222), and the Key Research Project of the National Natural Science Foundation of China (91531302).*

Crick, F.H.C. (1958). On protein synthesis. In Symposia of the Society for Experimental Biology, Number XII: The Biological Replication of Macromolecules, F.K. Sanders. (Cambridge: Cambridge University Press), pp. 138–163.

Eldredge, N., and Gould, S. (1972). Punctuated equilibria: an alternative to phyletic gradualism. In Models in Paleobiology, T.J.M. Schopf. (San Francisco: Freeman Cooper), pp. 84.

Jiang, D. (2009). Information Theory & Coding (in Chinese). 3rd ed. (Beijing: University of Science and Technology of China Press).

Karlin, S., and Ladunga, I. (1994). Comparisons of eukaryotic genomic sequences.. Proc Natl Acad Sci USA 91, 12832–12836.

Luo, L., Lee, W., Jia, L., Ji, F., and Tsai, L. (1998). Statistical correlation of nucleotides in a DNA sequence. Phys Rev E 58, 861–871.

Schmidt-Nielsen, K. (1984). Scaling: Why Is Animal Size So Important? (London: Cambridge University Press).

Schrodinger, E. (1944). What Is Life? The Physical Aspect of the Living Cell. (London: Cambridge University Press).

Shannon, C. (1949). The Mathematical Theory of Communication. (Chicago: University of Illinois Press).

Song, K., Ren, J., Reinert, G., Deng, M., Waterman, M.S., and Sun, F. (2014). New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. Brief Bioinform 15, 343–353.

Xu, Z., and Hao, B. (2009). CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Res 37, W174–W178.

SUPPORTING INFORMATION

Figure S1 Patterns of the information evolutionary rate.

Figure S2 Number of punctuated pattern in virtual evolutionary process in tested genome data.

Table S1 Classification of evolution pattern base on ratio of evolutionary rate

Table S2 Web sources of genome data

Table S3 Taxonomic relations of the selected species

The supporting information is available online at life.scichina.com and www.springerlink.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.