中国科学 生命科学
SCIENTIA SINICA Vitae

基因组演化的一个自然通信系统模型

吴琦, 王亚迪, 丁赟, 马帅, 吴宗敏 and 魏辅文

Citation: 中国科学: 生命科学 **47**, 1124 (2017); doi: 10.1360/N052017-00021

View online: http://engine.scichina.com/doi/10.1360/N052017-00021

View Table of Contents:http://engine.scichina.com/publisher/scp/journal/SSV/47/10

Published by the 《中国科学》杂志社

中国科学: 生命科学 2017 年 第 47 卷 第 10 期: 1124 ~ 1126
SCIENTIA SINICA Vitae lifecn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

点 评 中国科学院 科学与技术前沿论坛 后基因组时代的物理生物学

基因组演化的一个自然通信系统模型

吴琦[1], 王亚迪[2], 丁赟[1,3], 马帅[1,4], 吴宗敏[2*], 魏辅文[1*]

1. 中国科学院动物研究所, 动物生态与保护生物学院级重点实验室, 北京 100101;
2. 复旦大学数学科学院, 现代应用数学上海市重点实验室, 上海 200433;
3. 南昌大学生命科学学院, 南昌 330031;
4. 中国科学院大学生命科学学院, 北京 100049

* 联系人, E-mail: zmwu@fudan.edu.cn; weifw@ioz.ac.cn

收稿日期: 2016-12-06; 接受日期: 2016-12-22; 网络版发表日期: 2017-03-01

分子生物学的中心法则描述了生命现象中遗传信息的流动过程[1]. 但是, 这种“遗传信息”与物理学和信息与编码理论中的“信息”[2]意义颇为不同. 虽然物理学家很早就认为生命是一个从环境中吸收负熵的过程[3], 而负熵就是信息[2]. 但是迄今为止还没有一个成型的、物理的模型从信息传输的角度来描述生命进化的过程. 另一方面, 在生物学的研究中, 研究者通常倾向于关注生物性状的变异及其在时间中的进化, 而很少考虑性状与时间本身之间的联系. 换而言之, 进化生物学对性状变化的理解完全是牛顿的经典力学的方式. 本文尝试以时间和信息为基本量建立一个基因组进化的框架, 并试图讨论是否进化所产生的信息改变与进化所经历的时间之间会存在某种相互联系.

现有的通信系统涉及的是空间上不同的两个地点, 以及两点之间通过一定的物质性媒介的联系在信宿重建信源的统计状态的过程[2]. 与此相比, 生物的进化, 是时间上有差距的两个物种后者继承前者的遗传信息并发生部分改变的过程. 研究者们很容易将祖先物种类比为信源, 而新形成的后代物种类比为信宿. 但是这样的考虑将在信道问题上遇到困难: 很难设想一种存在于祖先和后代生命单元之间的可以供信息从中通过的通道. 本文尝试通过在信道概念中引入时间来解决这个问题. 不再把进化中的信道理解为一种占据空间的、“物质性”的东西, 而是理解为一种占据时间的、“事件性”的东西. 即把新物种从祖先物种中的形成过程中基因组的复制(或伴随着变异产生的新信息)看做一个事件, 这个事件本身被视为两代生命单元之间的唯一联系. 信道时间即为信源通过信道到达信宿的时间. 这样一来, 适合进化过程的自然通信系统, 就可以描述为物质性的信源和信宿通过一个事件性的信道联系起来并传递信息的过程. 根本上说就是祖先物种通过遗传信息的复制(和变异)形成后代物种, 这其中包括了进化革新(图1).

在人工通信系统中, 信源符号在进入信道前需要编码为可以在信道中直接传输的单义可译码再进入信道传输[4], 而信源编码存在一个最高编码效率[4]. 在自然通信系统中, 有类似的情况. 众所周知, 生命的演化过程是在选择压力下进行的, 因而不同种类信息传输系统的演化也同样应该是一个受自然选择影响的过程. 如果一个信息传输系统有较大的选择优势, 它就更有可能在漫长的生命历史中被保留下来. 可以想像, 一个信息传输系统如果有更高的信源编码效率和

引用格式: 吴琦, 王亚迪, 丁赟, 等. 基因组演化的一个自然通信系统模型. 中国科学: 生命科学, 2017, 47: 1124–1126, doi: 10.1360/N052017-00021
英文版见: Wu Q, Wang Y D, Ding Y, et al. A natural communication system on genome evolution. Sci China Life Sci, 2017, 60, 432–435, doi:10.1007/s11427–016–9011–7

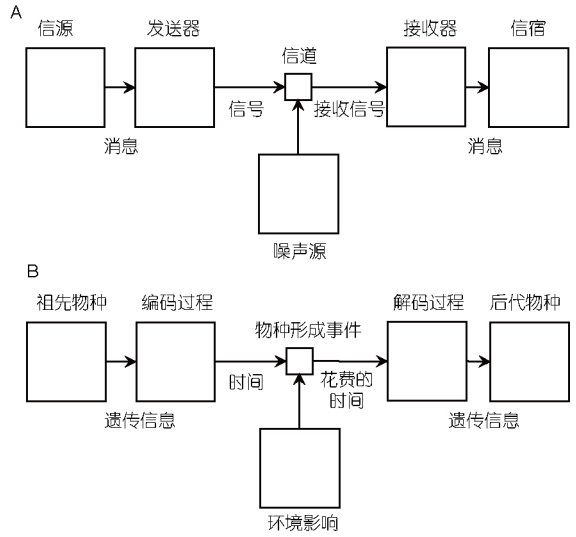

图 1 人工通信系统和自然通信系统的比较

A: 基于Shannon理论的经典人工通信系统示意图[2]; B: 自然通信系统. 遗传信息的编码和传递都是全基因组水平

符号发送速度, 那么在与其他信息传输系统的竞争中它会有较大的选择优势. 所以, 经历漫长的进化历史存留下来的现存生命过程中的信息通信系统一定会保持其信源编码效率非常接近其系统允许的最高效率, 称这一假定为自然通信系统的高效性假定.

在描述单义可译码的码符号在信道中传输所需的时间时, 通常会使用这样的表述方式“无噪音离散信道传输一个码符号需t秒时间”. 不妨把时间的持续称为“久度”. 分析这句话背后所预设的理论框架: “秒”这个时间单位自然的定义是来自“天”(一天有24小时, 每小时有3600秒), 也就是地球自转的久度. 所以这里隐含的一个关系是, 无噪音离散信道传输一个单义可译码符号的久度, 被用地球自转的久度来度量了. 当然可以换用其他的时间单位确定方式(如现在天文学中的秒是通过铯-133原子基态的能级跃迁周期来定义的), 但无论如何, 都在用一种空间中的运动来度量时间的久度以建立一个时间的单位, 并用这一时间单位来度量信息传输中时间的久度. 时间、空间和信息都是最基本的物理量. 上面这个预设框架显示, 在讨论信息与时间的关系时, 总是在引入空间来度量时间, 再用这个用空间作单位的时间来描述信息. 在自然通信系统中, 尝试去掉空间而直接用单义可译码的平均码长来度量信道中时间的久度. 回到上面举的例子, 不再说“以地球自转一圈所需时间久度的1/(24×3600)为单位, 离散无噪信道中传递一个码符号需要t个这样的单位”; 而是说“以码符号通过离散无噪信道所需的平均时间久度为一个单位, 信道中传递某一个码符号所需时间的久度是t个这样的单位”. 这样, 就在信息与时间的关系中排除了空间. 这意味可以将单义可译码码长理解为信源符号i在编码后经过信道到达信宿的时间来进行信源编码, 即有

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

其中k为常系数, 对应不同的信源编码效率, p_i为符号的信源概率, r为信源符号总数. 从中解出时间为

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

此式即为给定信源符号在最高效的编码后通过信道到达信宿的时间与该符号在信源的概率之间的关系. 取到等号时, 编码效率最高.

与人工通信系统的一个重要差别是, 在自然通信系统中信源和信宿的信息量会有差距. 因而给定符号i的信源概率p_s和信宿概率p_D会不同. 那么, 是哪个概率决定了信道的时间, 或者还有其他的概率描述方式来确定信道时间呢? 基因组序列由A, C, G, T 4种核苷酸组成, 所以基因组序列可以看作一个符号数为4的信源. 4种核苷酸的频率就描述了基因组的统计特征. 但是考虑到实际基因组序列的复杂性, 更符合实际的做法是考虑长度为k的序列中双核苷酸i, j的联合概率. 尽管定义不同, 相似的方法已用在量化基因组DNA序列的异质性中[5], 以此建立系统发生关系[6,7]或描述基因组进化机制[8]. 本文使用序列的互信息作为测度. 互信息的定义公式如下:

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

同时满足如下条件:

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

其中, i, j, k是A, C, G, T 4种核苷酸, 这里分别用1, 2, 3, 4表示(网络版附件1.1). 互信息公式实际上是一个十

维空间中的九维超曲面. 把这个曲面命名为“互信息曲面”. 一个给定双核苷酸频率分布的基因组对应着该空间曲面上的一个点. 因而, 信源向信宿的信息传输过程(或者说信源向信宿进化的过程)就可以看成互信息曲面中的一个点由坐标点S向坐标点D运动的过程. 由于前面讨论过的高效性假定, 这个运动一定会在这个九维曲面上找最短距离进行. 因而信息量曲面上信源到信宿的测地线长度即可作为信源到信宿的“信道距离”. 测地线弧长是一个概率空间中的无量纲量. 由式(3)知这个弧长决定了信道时间, 也即是祖先基因组通过自我复制产生变异的后代基因组的时间. 其可以表示为:

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

这里, k是常系数, L是弧长. 其微分形式为:

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

用信源与信宿的信息量之差作为基因组信息量的变化量, 用信源信息量通过信道的时间作为时间, 这样就可以定义信息量的进化速率μ如下式:

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

其中I_S是信源信息量, I_D是信宿信息量, $t_{S\to D}$是信息从信源点S到信宿点D通过信道所花费的时间, k, r和L的含义如前式(6). 其微分形式为:

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

式(8)显示, 进化速率μ由进化过程在测地线上的位置L, 信息增加量$\mathrm{d}I$和测地线弧长$\mathrm{d}L$共同决定.

简单地定性分析一下, 由于微分形式中有测地线长度项, 所以从祖先物种进化到后代物种的过程, 进化速率是逐渐增加的. 如果从一个祖先物种进化为两个后代物种, 两个后代物种将各有各的进化速率. 后代物种间的互信息量差异越大, 两条测地线弧长差异越大, 意味着各自的进化时间的差异就越大, 因而各自的进化速率的差异也就越大[9].

本研究组收集了94个来自公共数据库的基因组(网络版附表1和2), 涵盖了多细胞动物, 绿色植物和一些多细胞真菌还有单细胞真核生物. 然后用本模型计算了这些基因组数据的进化速率. 由于祖先基因组的双核苷酸频率难以推测, 依次将每个物种作为祖先其他所有93个物种作为后代来计算进化速率. 结果中, 有接近20%的进化过程起点和终点之间有大于一个数量级的差异(网络版附图1和2; 网络版附表3). 当选择某一个物种作为祖先时, 其他后代物种之间的系统发育关系越远, 得到的曲线的终点在粗粒平衡上的差距越大. 这些结果显示了某种间断平衡的进化模式[10].

参考文献

1 Crick F H C. On protein synthesis. In Sanders F K, ed. Symposia of the Society for Experimental Biology, Number XII: The Biological Replication of Macromolecules. Cambridge: Cambridge University Press, 1958. 138–163
2 Shannon C E. The Mathematical Theory of Communication. Illinois: University of Illinois Press, 1949
3 Schrodinger E. What Is Life? the Physical Aspect of the Living Cell. Cambridge: Cambridge University Press, 1944
4 Jiang D. Chinese, Information Theory & Coding. 3rd ed. Hefei: University of Science and Technology of China Press, 2009
5 Karlin S, Ladunga I. Comparisons of eukaryotic genomic sequences. Proc Natl Acad Sci USA, 1994, 91: 12832–12836
6 Song K, Ren J, Reinert G, et al. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. Brief Bioinform, 2014, 15: 343–353
7 Xu Z, Hao B. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucleic Acids Res, 2009, 37: W174–W178
8 Luo L, Lee W, Jia L, et al. Statistical correlation of nucleotides in a DNA sequence. Phys Rev E, 1998, 58: 861–871
9 Schmidt-Nielsen K. Scaling: Why Is Animal Size So Important? Cambridge: Cambridge University Press, 1984
10 Eldredge N, Gould S J. Punctuated equilibria: an alternative to phyletic gradualism. In: Schopf T J M, ed. Models in Paleobiology. San Francisco: Freeman Cooper, 1972. 84