文章编号：1007－4627(2007)02－0142－05

单峰高斯分布适应面上的误差阈*

冯晓利，李玉晓

（郑州大学物理工程学院，河南 郑州 450052）

摘 要：在 Eigen 的单峰适应面模型基础上，提出了生物体的适应值为高斯分布的随机适应面模型。利用系综平均的方法，计算了在单峰高斯分布适应面上准物种的浓度分布和误差阈。结果表明，对于小的适应面涨落，准物种分布和误差阈与确定情形相比变化极小，误差阈对于小的涨落是稳定的。然而，当适应值涨落较大时，从准物种到误差灾变的转变不再明显。误差阈变宽，并且在涨落增加时向大的突变率方向移动。

关 键 词：准物种；误差阈；高斯分布适应面

中图分类号：Q61 **文献标识码**：A

1 引言

自 Eigen[1] 提出自复制分子模型以来，对准物种演化的研究引起了人们的极大关注。在 Eigen 模型中，复制个体由化学容器中的生物大分子序列来表示；假定存在一种恒定的流来提供原料并且移去反应产物，以保持体系总浓度恒定；各分子类型的适应值由它在体系内单位时间所复制后代的数目（复制速率）来表示；复制过程中只存在点突变，并且各个复制位点具有相同的突变概率。模型中，突变和选择是两个相联系的过程，因此 Eigen 模型又称为耦合的突变选择模型。另外一种由 Crow 和 Kimura[2] 建立的模型，认为突变和选择是两个平行的独立过程。该模型也得到了广泛的研究[3, 4]。

目前，大量的研究工作主要集中于静态适应面[5, 6]。在简单的单峰适应面上，Eigen 模型得到两个显著的特征：当突变率足够小时，群体形成由突变型序列紧密围绕在主序列（野生型）周围的一种分布，即“准物种”；当突变率超过某个临界值时，群体就散布在整个基因型空间，每种分子序列等概率出现，该转变点即所谓的“误差阈”。这两个特征在许多不同适应面上[3, 7－11]也得到了证实。误差阈特性已经用来理解真实生物体的演化[12, 13]，对误差转变理论的应用也有一定探讨。最近几年，对于时间依赖适应函数和动态适应面的情形也有一定研究[14—16]，并建立了一套能够在周期适应面上定义和数值计算准物种的方法。事实上，实际的生物体系常常受到外界环境变化的影响，而这种变化在很多情况下是无法控制和难以预测的，通常是随机的。因此，在某种程度上，生物体的适应值也呈现一定随机行为。

为说明适应面的涨落（环境噪声）对生物分子平衡浓度分布及误差阈的影响，本文中假定适应值服从高斯随机分布。从统计的观点，我们给出并分析了系综平均的准物种分布和误差阈的数值模拟结果。我们更为关注的是平均误差阈的一般特征。

2 模型与公式

生物体产生后代的过程可以看作是一定核苷酸序列的宏观大分子的复制过程。正如在 Eigen 模型中，体系的分子是由长度为 n 的字符序列表示，每个字符可取 κ 类不同的字符。在 DNA 或 RNA 结构中，κ 取 4 种不同类型（G，A，C，T/U），为简单起见，这里 κ 取 2，仅仅区分嘌呤（R）和嘧啶（Y）。我们考虑突变完全来自碱基替换（点突变），则可能的序列总的数目为 $N = 2^n$。假定群体数目很大并且保持恒定。我们用 S_i（$i = 1, 2, \cdots\cdots 2^n$）标记第 i 种类型序列，用 x_i 表示序列 S_i 的浓度，则突变选择方程可写为

* **收稿日期**：2007－01－08；**修改日期**：2007－03－20

作者简介：冯晓利（1981－），女（汉族），河南汝州人，博士生，从事统计物理、理论生物物理研究；E-mail：xlf32@163.com

$$\frac{\mathrm{d}x_i}{\mathrm{d}t}=\sum_{k=0}^{N}W_{ik}x_k-[D_i+\Phi(t)]x_i, \quad (1)$$

其中 W_{ik} 是复制速率矩阵元素，可写为

$$W_{ik}=A_kQ_{ik}, \quad (2)$$

这里 A_k 是第 k 种类型的分子复制速率，(Q_{ik})表示突变矩阵，Q_{ii} 表示完全正确复制 S_i 的概率，Q_{ik} 表示由于复制 S_k 而得到 S_i 的概率。单个位点的复制精度 q ($0\leqslant q\leqslant 1$)，表示正确复制一个字符的概率，并且假定 q 对序列中所有位置都是相同的，则突变矩阵给出：

$$Q_{ii}=q^n, \quad (3)$$

$$Q_{ik}=q^{n-d(i,k)}(1-q)^{d(i,k)} \quad (i\neq k), \quad (4)$$

其中 $d(i,k)$ 是序列 S_i 和 S_k 之间的 Hamming 距离，它表示序列 S_i 和 S_k 所对应的不同字符的数目。D_i 表示序列 S_i 的死亡率，为简单起见，假定 $D_i=D$。由于对每种分子序列，相同的 D 值使得总体系的平均适应值产生一定平移，而并不改变群体的相对浓度分布。为此，我们取 $D=0$。$\Phi(t)$ 是为保持总浓度不变而引入的稀释流，可由条件 $\sum_i \mathrm{d}x_i/\mathrm{d}t=0$ 确定。

方程(1)是非线性的，可通过适当的线性变换去掉非线性项而直接得其解[17]。引入变量 $z(t)$：

$$z(t)=\exp\left(\int_0^t\phi(\tau)\mathrm{d}\tau\right)x(t), \quad (5)$$

将方程(5)代入方程(1)得到线性方程：

$$\frac{\mathrm{d}z_i}{\mathrm{d}t}=\sum_{k=0}^{N}W_{ik}z_k, \quad (6)$$

W 矩阵的所有元素都是正的，其最大本征值对应的归一化本征矢即为稳态时不同序列的相对浓度。由于 $x(t)$ 和 $z(t)$ 只差了一个标量因子，于是 $x(t)$ 可表示为

$$x_i(t)=\frac{z_i(t)}{\sum_{k=0}^{N}z_k(t)}, \quad (7)$$

矩阵 Q 的维数随 n 的增加成指数增加，这使得数值解和解析解都很难处理。正如前面所述，我们按照所有序列与主序列的 Hamming 距离将其归类，将 Hamming 距离为 i 的所有序列总和归为突变类 I_i($i=0,1,\cdots\cdots n$)。这样，所有序列可分为 $n+1$ 类。我们假定同一突变类的序列具有相同的适应值 A_i，并用 y_i 表示突变类 I_i 的相对浓度，则方程化简为

$$\frac{\mathrm{d}y_i}{\mathrm{d}t}=\sum_{k=0}^{n}A_kQ'_{ik}y_k-fy_i, \quad (8)$$

相应于 $\Phi(t)$，f 的引入是为了保持总浓度恒定，于是 $f(t)=\sum_{k=0}^{n}A_ky_k$。$Q'_{ik}$ 是从 I_k 类到 I_i 类的突变概率，其表达式为

$$Q'_{ik}=\sum_{l=l_{\min}}^{l_{\max}}\binom{k}{l}\binom{n-k}{i-l}q^{n-k-i+2l}(1-q)^{k+i-2l}, \quad (9)$$

其中 $l_{\min}=\max\{0,k+i-n\}$，$l_{\max}=\min\{k,i\}$。

在我们的模型中，所有序列的适应值都假定为高斯随机变量。对于主序列，其概率分布为

$$P(A_0)=\frac{1}{\sqrt{2\pi\omega_0^2}}\exp\left[-\frac{(A_0-a)^2}{2\omega_0^2}\right], \quad (13)$$

其中平均适应值为 a_0，方差为 ω_0^2。对每一类突变体，满足以下相同的概率分布：

$$P(A_i)=\frac{1}{\sqrt{2\pi\omega_1^2}}\exp\left[-\frac{(A_i-a_1)^2}{2\omega_1^2}\right], \quad (14)$$

其中平均适应值为 a_1，方差为 ω_1^2。模型中，假定主序列的平均适应值比突变体的平均适应值大。不失一般性，并便于和单峰适应面模型情况进行比较，取 $a_0=10$，$a_1=1$。

3　结果与讨论

在确定的单峰适应面模型中，序列长度与误差阈成反比的关系[10]。在数值模拟中，为便于比较，序列长度取为 $n=20$。对给定适应面的一个实现，每一个突变类序列获得一个不同的随机适应值。我们进行了 1 000 或 10 000 次随机抽样并进行了系综平均。适应值涨落的强度可由 ω_0/a_0 和 ω_1/a_1 表示。为了得到准物种和误差阈在高斯随机分布适应面上的变化，我们从小到大地改变 ω_0 和 ω_1 的值，计算了平均准物种分布和平均误差阈。

为此，首先给出在确定适应面情形下不同类型分子的相对浓度。在图 1 中，适应面为单峰适应面。数字 0 代表主序列，数字 1，2，3，…分别代表突变类 1，2，3，…。很明显，误差阈位于 $1-q\approx 0.11$。

我们感兴趣的是平均相对浓度和平均误差阈随

适应面涨落强度的不同如何变化。数值模拟表明，当分布宽度较小时，平均浓度和误差阈与确定情形相比变化很小。这可在图 2 中看到，其中分布宽度为 $\omega_0 = 0.1$ 和 $\omega_1 = 0.01$，其余参数保持不变(在这种情形下 $\omega_0/a_0 = \omega_1/a_1 = 0.01$)。这表明浓度分布和误差阈对小的外界涨落是稳定的，临界转变点在误差阈处是稳定的。

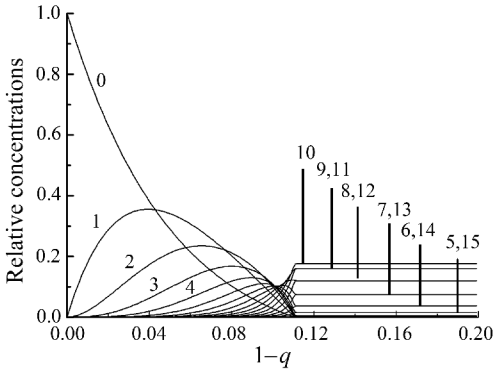

图 1 在单峰适应面上，不同突变类在平衡态的相对浓度分布随单点错误率 $(1-q)$的变化关系
参数：$n=20$，$a_0=10$，$a_1=1$。

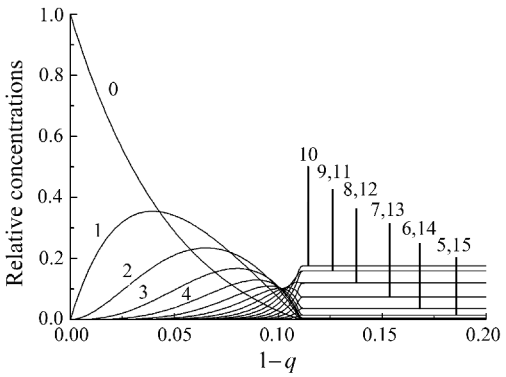

图 2 在单峰高斯分布适应面上，不同突变类在平衡态的相对浓度分布随单点错误率 $(1-q)$的变化关系
参数：$a_0=10$，$\omega_0=0.1$，$a_1=1$，$\omega_1=0.01$。

当分布宽度增大到与平均适应值 a_0 和 a_1 可相比的值时，群体浓度从准物种分布到随机分布的转变点不再明显，平均误差阈变得平滑并且向较大值有一个明显的平移。这可由图 3 说明，其中 $\omega_0/a_0 = \omega_1/a_1 = 0.25$。当涨落再增大时，体系变得不稳定。由于分布宽度太大将会出现负的适应值和负的群体浓度，这对我们的研究没有意义。利用截尾正态分布，得到的结果与图 3 类似。

为了更清楚地看到高斯分布宽度对浓度分布和误差阈的影响，图 4 给出了单个突变类($I_i=10$)序列在不同分布宽度下的浓度曲线。可以看到，与确定情形相比，只有当分布宽度较大时，平均浓度在误差阈附近的变化才比较明显，它随着突变率的增加缓慢地达到最大值。而在远离误差阈的区域，平均浓度和确定情况下的浓度几乎没有差别。这表明在误差阈区域以外，浓度分布对于适应值涨落仍然是稳定的。

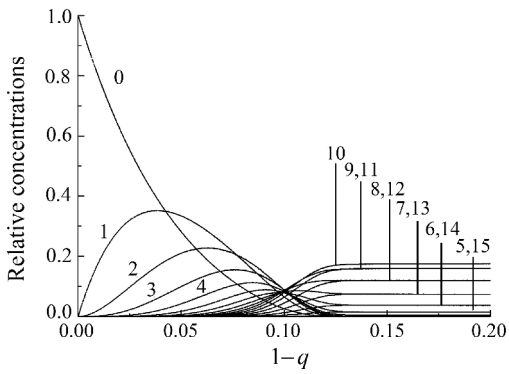

图 3 在单峰高斯分布适应面上，不同突变类在平衡态的相对浓度分布随单点错误率 $(1-q)$的变化关系
参数：$a_0=10$，$\omega_0=2.5$，$a_1=1$，$\omega_1=0.25$。

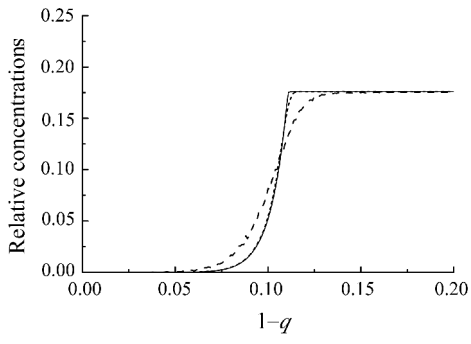

图 4 在两种不同适应面上，与主序列的 Hamming 距离为 10 的突变类在平衡态的相对浓度分布
参数：$n=20$；— $a_0=10$，$a_1=1$；┈ $a_0=10$，$a_1=1$，$\omega_0=0.5$，$\omega_1=0.05$；┈ $a_0=10$，$a_1=1$，$\omega_0=2.5$，$\omega_1=0.25$。

为了理解误差阈作为相变的性质，并更好地比较浓度分布和误差阈在随机的和确定的适应面情形下的不同之处，我们引入序参数 m。另外一种表示方法可由群体与主序列之间的平均 Hamming 距离 $\langle d\rangle$给出，这里$\langle\rangle$表示在平衡态对群体的统计平均。它们之间的关系为 $m=1-2\langle d\rangle/n$ 。在 $m=1$，群体完全由主序列构成；在 $m=0$，群体变得完全无序，每个序列以相同的概率出现；在 $m=-1$，群体

完全由主序列 S_0 的互补序列 S_n 构成。

图 5 给出了序参数作为单点突变率的函数曲线。对于小的分布宽度，序参数几乎与确定的情形重合；对于大的分布宽度，序参数在误差阈处连续地变为 0，表明适应值的涨落对误差阈有较大的改变。因此，相变只有在涨落较小时才存在，涨落较大时，相变消失。

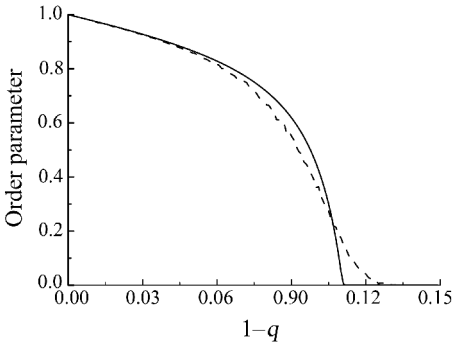

图 5　在两种不同适应面上，序参数在平衡态下的分布
参数：$n=20$，$a_0=10$，$a_1=1$；— 表示确定适应面情形，--- 表示高斯随机适应面情形：$\omega_0=2.5$，$\omega_0=0.25$。

按照对误差阈普遍认同的定义，其最显著的特征是在平衡态野生型(主序列)的浓度变为 0。通过对 Eigen 方程在稳态解的分析可知，除非在 $n\to\infty$ 的情况[18]，否则主序列的浓度并不真正为 0。因而，具有相变特性的误差阈只有在 $n\to\infty$ 时才存在。事实上，任何生物体的核苷酸序列长度都是有限的。在本文中，取其长度为 20。很明显，对于确定的适应面，并不存在这样一种等同于相变的误差阈，误差阈应该处于一个很小的范围之内。对我们的单峰高斯分布适应面，误差阈区间随着分布宽度的增加而变宽。

4　结论

由于实际生物体演化的复杂性，特别是外界许多不确定性因素的影响，我们很难确定其适应面的函数形式。在 Eigen 单峰适应面基础上，通过引入高斯分布的随机适应面，我们计算了它的浓度分布和误差阈，并和确定适应面模型的结果进行了比较。结果表明，小的适应面涨落对误差阈的影响非常小，误差阈在小的适应值变动下是稳定的。然而在相当大的适应面涨落时，误差阈变得平滑并向较大值有一定平移，误差阈处于一定的范围之内。

致谢　感谢中国原子能科学研究院核物理研究所的卓益忠研究员和顾建中教授的指导和有益的讨论。

参考文献(References)：

[1] Eigen M. Naturwissenchaften, 1971, **58**: 465.
[2] Crow J F, Kimura M. An Introduction to Population Genetics Theory. New York: Harper and Row, 1970, 656.
[3] Baake E, Baake M, Wagner H. Phys Rev Lett, 1997, **78**: 559.
[4] Wiehe T, Baake E, Schuster P. J Theor Biol, 1995, **177**: 1.
[5] Swetina J, Schuster P. Biophys Chem, 1982, **16**: 329.
[6] Galluccio S. Phys Rev, 1997, **E56**: 4 526.
[7] Woodcock G, Higgs P G. J Theor Biol, 1996, **179**: 61.
[8] Tarazona P. Phys Rev, 1992, **A45**: 6 038.
[9] Paulo R A Campos. Phys Rev, 2002, **E66**: 062 904.
[10] Eigen M, McCaskill J, Schuster P. Adv Chem Phys, 1989, **75**: 149.
[11] Leuthausser I. J Stat Phys, 1987, **48**: 343.
[12] Nilsson M, Snoad N. Phys Rev Lett, 2000, **84**: 191.
[13] Kamp C. Microbes and Infection, 2003, **5**: 1 397.
[14] Wilke C, Ronnewinkel C, Martinetz T. Phys Rep, 2001, **349**: 395.
[15] Nilsson M, Snoad N. Phys Rev, 2002, **E65**: 031 901.
[16] Kamp C, Bornholdt S. Phys Rev Lett, 2002, **88**: 068 104.
[17] Thompson C J, McBride J L. Math Biosci, 1974, **21**: 127.
[18] Eigen M. J Biophys Chem, 2000, **85**: 101.

Error Threshold on Single Peak Gaussian Distributed Fitness Landscapes*

FENG Xiao-li[1), LI Yu-xiao

(*School of Physical Engineering, Zhengzhou University, Zhengzhou* 450052, *China*)

Abstract: Based on the Eigen model with a single peak fitness landscape, the fitness values of all sequence types are assumed to be random with Gaussian distribution. By ensemble average method, the concentration distribution and error threshold of quasispecies on single peak Gaussian distributed fitness landscapes were evaluated. It is shown that the concentration distribution and error threshold change little in comparing with deterministic case for small fluctuations, which implies that the error threshold is stable against small perturbation. However, as the fluctuation increases, the situation is quite different. The transition from quasi-species to error catastrophe is no longer sharp. The error threshold becomes a narrow band which broadens and shifts toward large values of error rate with increasing fluctuation.

Key words: quasispecies; error threshold; Gaussian distributed fitness landscape

* **Received date**: 8 Jan. 2007; **Revised date**: 20 Mar. 2007

1) E-mail: xlf32@163.com