文章编号： 1007-4627(2016) 04-0500-06

基于 Logistic 回归建模和马尔可夫链蒙特卡罗方法计算后验描述丁酸梭菌株对于给定辐照剂量区的应答趋势

周 翔[1]，姜婷婷[1,2]，徐 丹[1]，杨 榛[1,3]，梁剑平[1]，王 亮[1,2]

(1. 中国科学院近代物理研究所，兰州 730000；
2. 中国科学院大学，北京 100049；
3. 南京农业大学，南京 210095)

摘要： 利用马尔可夫链蒙特卡罗 (Markov Chain Monte Carlo, MCMC) 方法估计 Logistic 回归模型中的参数，就是要构造一个以参数的后验分布为其平稳分布的非周期不可约的马尔可夫链，然后用该平稳分布中抽出的样本点计算蒙特卡罗积分。上述理论方法可以解决实验样本数据由于存在定和约束和多重共线性、在进行经典的 logistic 回归建模时的困难问题。基于此方法, 研究了丁酸梭菌株对于给定辐照区间剂量的应答趋势，用模型挖掘数据所隐含的内在信息并导出了 Logistic 回归模型参数的贝叶斯框架下的 50%，90%，95% 和 99%的置信区间。结果表明，运用 Logistic 与马尔可夫链耦合模型在有关给定辐射剂量对于微生物作用效果问题的 logistic 回归建模中具有较大的科学性与很好的使用性，从而可以为辐照诱变处理微生物制定辐照剂量区提供理论支持和回归技术借鉴。

关键词： Logistic回归；马尔可夫链；辐射剂量；丁酸梭菌

中图分类号： R815.2；TL72　　**文献标志码：** A　　**DOI：** 10.11804/NuclPhysRev.33.04.500

1 引言

线性回归模型 (Linear Regression Model) 的应用相当广泛，已经渗透到医学、经济学、生物学、犯罪心理学、工程技术学等领域主要原因是回归模型是处理分类数据的有力工具，且对解释变量几乎没有任何限制[1−2]。建立回归模型与其他传统模型一样，主要有两个目的：(1) 用模型去挖掘数据所隐含的内在信息，以及用模型去衡量解释变量与响应变量的相依关系；(2) 预测或为决策者提供某些先验信息，作出较准确的决策[3−4]。然而，线性回归分析一般要求响应变量是连续变量、数据分布为正态分布等条件。在实际分析研究中，经常遇到的是非连续的响应变量，即分类响应变量。在研究二分变量与诸多自变量的相互关系时，常选用 Logistic 回归 (Logistic Regression) 模型[5]。Logistic 回归是研究因变量为二分类观察结果与影响因素 (自变量) 之间关系的一种多变量分析方法，属概率型非线性回归[6−7]。

非线性回归模型都是复杂的，其参数的估计也必然要应用马尔可夫链蒙特卡罗 (Markov Chain Monte Carlo，MCMC) 方法才能得以实现[8−9]。随着计算机技术的发展和贝叶斯方法的改进，特别是 MCMC 方法的发展，使原先复杂的高维计算问题迎刃而解，很大程度上方便了参数的后验推断问题[10−12]。利用 MCMC 方法估计 Logistic 回归模型中的参数，就是要构造一个以参数的后验分布(目标分布)为其平稳分布的非周期不可约的马尔可夫链，然后用该平稳分布中抽出的样本点计算蒙特卡罗积分[13−16]。常见的抽样方法有：Importance Sampling[17]，Hammersley & Handseom[18]，Metropolis-Hasting Sampling[19]；MH，Metropolis[20] 以及 Gibbs[21] 抽样方法。其中 Metropolis-Hasting Sampling 抽样方法是 MCMC 方法的基本框架，也是重要抽样的一种实现。

近年来，随着世界不断增长的可再生燃料的需求，一种新生代生物能源——生物丁醇正在进入人们的视野。生物丁醇在燃料性能和经济性方面具有明显的

收稿日期： 2016-03-21；　**修改日期：** 2016-05-25

基金项目： 中国科学院西部之光人才培养引进计划——“西部青年学者”A类项目(Ke-Fa-Ren-Zi[2015] No.77)；甘肃省自然科学基金项目(1506RJZA293)

作者简介： 周翔(1977–)，男，兰州人，博士后，副研究员，硕士研究生导师，从事生物物理与代谢工程研究；
E-mail：syannovich@gmail.com；syannovich@impcas.ac.cn。

优势，比乙醇有着更好的应用前景。丙酮、乙醇、丁醇 (acetone-ethanol-butanol，简称 ABE) 发酵是一项传统的大宗发酵丁醇工业，而丁酸梭菌 (*Clostridium tyrobutyricum*) 协同丙酮-丁醇梭菌 (*Clostridium aceto-butylicum*) 可产生大量的丙酮、丁醇和乙醇等溶剂，是重要的工业发酵丁醇的菌种之一[22−23]。目前利用重离子束辐照诱变处理微生物菌株，为行业提供高转化率的发酵新菌种并使之产业化，已显示出极大的优势和先进性，开创了重离子在微生物育种上应用的崭新局面[24−25]。描述给定辐照剂量对于该菌株作用效果的研究是当前逆代谢工程与微生物发酵工业科学的热点领域，开展重离子束辐照诱变影响下该菌株不同应答变化趋势，可为制定科学、有效的重离子束辐照计划提供决策，为诱变处理微生物格局优化提供依据。Logistic 回归建模和 MCMC 方法综合了 Logistic 模型模拟复杂系统变量与响应变量变化的能力和 Markov 模型长期预测的优势，既提高重离子束辐照诱变处理的预测精度，又可有效模拟该菌株不同应答变化趋势，有较大科学性与使用性。因此，本文选择重离子束辐照诱变丁酸梭菌为研究内容，运用 Logistic 与 Markov 耦合模型并设置可能的菌株应答变化趋势情景，用模型挖掘数据所隐含的内在信息，以期为该菌辐照诱变处理制定辐射剂量提供支持和借鉴。

2　理论与方法

2.1　Logistic模型与Logistic回归模型

基于二项分布族的广义线性模—Logistic 回归模型是研究因变量为二分类观察结果与影响因素 (自变量) 之间关系的一种多变量分析方法，属概率型非线性回归。而响应变量Y是二分类变量，取值为 1 和 0。常用的 Logistic 模型为

$$P(Y=1|x_1,x_2\cdots,x_n)=\frac{\exp(\beta_0+\beta_1x_1+\cdots+\beta_nx_n)}{1+\exp(\beta_0+\beta_1x_1+\cdots+\beta_nx_n)}\ , \tag{1}$$

其中：x_1, …, x_n 为回归模型的解释变量，β_0, β_1, ⋯, β_p 为似然方程的解。

对式(1)作 logit 变换，Logistic 回归模型可以写成：

$$\text{logit}(p)=\ln\left[\frac{p(Y=1)}{1-p(Y=1)}\right]=\beta_0+\beta_1x_1+\cdots+\beta_nx_n+\varepsilon\ , \tag{2}$$

其中 x_1, ⋯ ,x_n 为回归模型的解释变量。误差项 ε 的分布与Y的分布有关。本文假定Y 和 ε 都服从伯努利分布。Logistic 回归参数主要使用最大似然法 (Maximum Likelihood Estimation) 估计。令容量为 n 的样本 Y_1, ⋯ ,Y_n，则似然函数为

$$L(Y_1,\cdots Y_{\rm n})=\prod_{i=1}^{n}P(Y_1=1)^{Y_i}(1-P(Y_1=1))^{1-Y_i}\ , \tag{3}$$

对数似然函数为

$$\begin{aligned}\ln[L(Y_i,\cdots,Y_n)]=&\sum_{i=1}^{n}\left[Y_i\ln(P(Y_i=1))+\right.\\&\left.(1-Y_i)\ln(1-P(Y_i=1))\right]\\=&\sum_{i=1}^{\rm n}\left[Y_i(\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip})\right]-\\&\sum_{i=1}^{n}\ln\left(1+{\rm e}^{\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}}\right)\ ,\end{aligned} \tag{4}$$

其中 $x_{i1},\cdots,x_{ip}$ 是与 Y_i 相对应解释变量的观测值。将对数似然函数分别对 β_0, β_1, ⋯ ,β_p 求偏导数，并令偏导数为 0，得到似然方程。似然方程的解 β_0, β_1,⋯ ,β_p 为回归参数 β_0, β_1, ⋯ ,β_p 的估计值。最大似然估计具有的一致性、有效性和正态性都是一些很好的统计性质，样本数据越大时其估计值就越准确。

2.2　MCMC的可靠度预测值的数值计算方法

为了计算未知的随机参数及误差项标准差的后验概率密度函数 $p(\theta,\sigma_\varepsilon|Y)$，首先假定随机参数 θ 和误差项标准差 σ_ε 的联合先验分布 $\pi(\theta,\sigma_\varepsilon)$，在已知 $\pi(\theta,\sigma_\varepsilon)$ 的基础上，得到观测数据集 y 后就可以利用贝叶斯公式来推断 $p(\theta,\sigma_\varepsilon|Y)$，即有

$$p(\theta,\sigma_\varepsilon|Y)=\frac{f(Y|\theta,\sigma_\varepsilon)\pi(\theta,\sigma_\varepsilon)}{\int_\Theta\int_{R^+}(Y|\theta,\sigma_\varepsilon)\pi(\theta,\sigma_\varepsilon)\,{\rm d}\sigma_\varepsilon{\rm d}\theta}\ , \tag{5}$$

其中，$f(Y|\theta,\sigma_\varepsilon)$ 是在参数为 θ、误差项标准差为σ_ε 的情况下观测数据集 y 的联合概率密度函数，也就是似然函数。根据误差项的正态独立性可知：

$$f(Y|\theta,\sigma_\varepsilon)=\prod_{i=1}^{\rm k}\phi\left[Y_i;\eta(dose_i;\theta),\sigma_\varepsilon^2\right]\ , \tag{6}$$

其中，$\phi\left(Y;\mu,\sigma^2\right)$ 表示均值为 μ、方差为 σ^2 的正态分布概率密度函数在 y 处的取值。在已知未知的参数及误差项标准差的后验分布之后，求取如下的后验期望可以

表示为下式：

$$E^{p(\theta,\sigma_\varepsilon|Y)}\left[\mathrm{g}(\theta,\sigma_\varepsilon;dose_k,\Delta_{\mathrm{dose}}),D\right]=\frac{\int_\Theta\int_{R^+}g(\theta,\sigma_\varepsilon;dose_k,\Delta_{\mathrm{dose}})f(Y|\theta,\sigma_\varepsilon)\pi(\theta,\sigma_\varepsilon)\,\mathrm{d}\sigma_\varepsilon\mathrm{d}\theta}{\int_\Theta\int_{R^+}f(Y|\theta,\sigma_\varepsilon)\pi(\theta,\sigma_\varepsilon)\,\mathrm{d}\sigma_\varepsilon\mathrm{d}\theta}。\tag{7}$$

一般说来，要得到如式(7)所示的后验期望的解析表达式是困难的，通常需要借助于数值方法或近似方法进行计算,如数值积分、蒙特卡罗积分或解析近似方法等。特别地，在这里σ_ε未知，式(6)所示的似然函数中包含分母上的随机变量σ_ε，并且σ_ε的先验分布又是无信息先验分布，从而后验分布将会非常复杂。为了解决这个问题，采用MCMC方法建立一个平稳分布为$p(\theta,\sigma_\varepsilon|Y)$的马尔可夫链，通过该马尔可夫链$p(\theta,\sigma_\varepsilon|Y)$的样本，从而得到计算后验期望。其具体实现方法，本研究采用单元素Metropolis-Hasting Sampling方法。

2.3 数据抽样采集

重离子辐照实验在兰州重离子研究装置(HIRFL)TL2终端上进行。重离子束流为$^{12}C^{6+}$离子束，240 *A*MeV，LET为35.5 keV/μm，吸收剂量率3 Gy/min，用空气电离室监测剂量。菌株细胞为随机抽取数。随机辐照随机抽取的酮丁醇梭菌细胞所用剂量分别为15～85 Gy。采用MTT法数据处理菌株细胞的存活数。随机抽取样本点，选取第6组样本作为可靠性分析及预测研究的对象，而其余40组样本用来提取有关随机参数的先验信息。

3 结果和讨论

对于每个菌株细胞，其结果是要么死亡(成功，1)，要么存活(失败，0)。为了生成随机数，定义函数$f(X_1,\ \cdots,\ X_p)=b_0+b_1X_1+\cdots+b_pX_p+\varepsilon$，其中$b_0,\ b_1,\ \cdots,\ b_p$是函数服从均值为0、方差为$\sigma^2$的正态分布，记为$\varepsilon\sim N(0,\ \sigma^2)$。$X_1,\ \cdots,\ X_p$是标准正态分布所生成的随机数系数(已知的)，由标准正态分布随机数代替。容量为n的样本生成过程如下。首先，生成$p+1$个标准正态分布的随机数，次令$b_0,\ b_1,\ \cdots,\ b_p$等于这$p+1$个数。其次，生成n组正态分布随机数$x_{i1},\ \cdots,x_{ip}(i=1,\ \cdots n)$，$n$个服从$N(0,\ \sigma^2)$的随机数赋给$\varepsilon_1,\ \cdots,\ \varepsilon_n$。计算函数值$f_i(x_{i1},\ \cdots,\ x_{ip})=b_0+b_1x_{i1}+\cdots+b_px_{ip}+\varepsilon_i$，并计算函数值的中位数，记为*Median(f)*。最后,对于$i=1,\ \cdots,n$，令：

$$y_i=\begin{cases}1, & f_i(x_{i1},\cdots,x_{ip})\geqslant Median(f)\\ 0, & f_i(x_{i1},\cdots,x_{ip})\prec Median(f)\end{cases},\tag{8}$$

则$y_i,\ x_{i1},\ \cdots,\ x_{ip}(i=1\cdots,\ n)$即为所分析的样本数据。当有多个函数值等于*Median(f)*时，需要继续生成一些随机数，以保证类别1和类别0的容量都等于$n/2$。辐射实验结果构成的总体，$Y_1,\ \cdots,\ Y_n$。从中随机抽取n个辐射结果作为样本，实验值标注为$y_1,\ \cdots,\ y_n$，设$p_i=P(y_1=1|x_i)$为给定辐射剂量x_i的条件下得到的结果$y_i=1$的条件概率：而同样条件下得到结果$y_i=0$的条件概率为$P(y_1-0|x_i)=1-p_i$。于是得到如下的实验概率：

$$P\{y(dose_k+\Delta dose)\leqslant D|\ y,y_i\leqslant D,i=1,\cdots,k\}\ ,\tag{9}$$

其中，$y(dose_k+\Delta dose)=\eta(dose_k+\Delta dose;\theta)+\varepsilon_{k+1}$；$dose_k$是辐射剂量；$y=\{y_1,\cdots,y_k\}$是给定辐射剂量对于菌株细胞的作用效果的数据集，$\Delta dose$是我们所感兴趣的辐射区间的剂量；$\varepsilon_{k+1}\sim N(0,\ {\sigma_\varepsilon}^2)$表示退化轨道模型在$dose_k+\Delta dose$给定剂量的随机误差项，设它同$\{\varepsilon_1\cdots,\varepsilon_k\}$独立。正如图1所示，无论$\varepsilon_k$取任何值，Logistic函数$p_i=P(y_1=1|x_i)=1/(1+e^{\varepsilon_k})$的取值范围在0和1之间。本文运用样本参数与能力参数交替迭代计算的方法生成马尔可夫链；然后当马尔可夫链达到平稳分布时, 采取迥然不同于极大似然方法的思路，用平稳分布中的抽样点来计算蒙特卡罗积分。当ε_k增大时，这一函数值先是缓慢地增加，然后迅速增加，之后增加的速度又开始逐渐减缓，最后当ε_k趋近于$+\infty$时，Logistic回归函数值趋近于1。Logistic回归模型S型曲线表明，在给定辐射剂量很小(ε_k值很小)时其对于该菌株细胞的致死作用也很小，然而在中间阶段对应的致死作用增加很快，但是在给定辐射剂量85 Gy以后(ε_k值很大)，对于该菌株细胞的致死作用就保持在几乎不变的水平了。值得注意的是，近年来生物高新技术的发展和应用，使低剂量辐射生物效应研究有了新的发现，特别是传统放射生物学理论无法完全解释的现象，如适应性反应(兴奋效应)、旁效应、超敏感反应等。这些效应的产生一方面与剂量水平有关，另一方面也与辐射的品质(不同LET)有关，如有关旁效应的绝大部分报告是α粒子辐射效应，而辐射兴奋效应几乎只出现在低LET辐射，如X或γ射线照射细胞或整体动物。因此，对于丁酸梭菌株在低剂量的辐射效应机制和评价值得深入研究。同样是低剂量辐射，但所表现出来的可能

是绝然不同的效应，如适应性反应与旁效应，这固然与辐射品质 (不同 LET) 和剂量有关，但其发生的机制如何，目前还不是很清楚，而且上面所提到的效应是暂时早期效应，有关其远后效应尚缺乏有效的检测和评价技术体系。

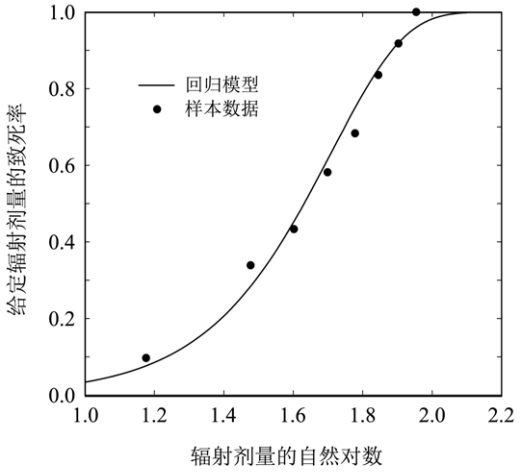

图 1　丁酸梭菌细胞死亡与给定辐射剂量的Logistic回归模型

为了预测Logistic 回归模型可靠度曲线，本文选取该样本作为可靠性分析及预测研究的对象，而其余40 组样本用来提取有关随机参数的先验信息。随机参数先验分布为独立正态分布，通过分析可以得到$\pi\theta(\theta)=\pi_1(\theta_1)\pi_2(\theta_2)$，$\pi_1(\theta_1)$和$\pi_2(\theta_2)$分别是$\theta_1$和$\theta_2$的先验概率密度函数，它们的具体分布如图2所示，可以看作被选取样本点θ_1和θ_2的先验概率密度。随后，根据参数向量$Z^{(n-1)}$产生$Z(n)$, $n=1, 2, \ldots$。

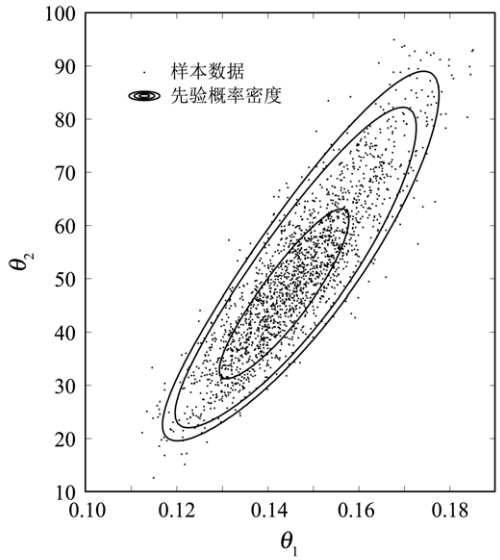

图 2　抽取样本数据分布与先验概率密度函数

考虑第i个分量Z_i, $i=1, \ldots, q+1$的转移，根据Z_i的建议分布产生$T_i\left(Z_i^{(n-1)}\to Z_i^{(n)}\middle| Z_{-i}{}^{(n-1)}\right)$一个可能的$Z_i^{(n)}$然后根据概率：

$$b_i\left(Z_i{}^{(n-1)}\to Z_i{}^{(n)}\middle| Z_{-i}{}^{(n-1)}\right)=\min\left\{1,\frac{P'\left(Z_i{}^{(n-1)}, Z_{-i}{}^{(n-1)}\middle| y\right)}{P'\left(Z_i{}^{(n-1)}, Z_{-i}{}^{(n-1)}\middle| y\right)}\right\}, \tag{10}$$

来决定是否转移，其中：

$$Z_i{}^{(n-1)}\left[Z_i{}^{(n)}\cdots Z_{i\text{-}1}{}^{(n)}Z_{i+1}{}^{(n-1)}\cdots Z_{\mathrm{q}+1}{}^{(n-1)}\right]。 \tag{11}$$

重复式(11)步直到Z各分量的遍历均值都稳定后，采集所需样本并终止马尔可夫链计算。

最终得到的马尔可夫链长为N (不计$Z^{(0)}$)，去掉遍历均值稳定之前的M个迭代值，将马尔可夫链中后面的$N\sim M$个迭代结果$Z_{M|1}\cdots, Z_N$用于下一步计算。

利用下式计算式(12)所示的后验期望：

$$E^{p(\theta,\sigma_\varepsilon|Y)}\left[\mathrm{g}\left(\theta,\sigma_\varepsilon;dose_k,\Delta_{dose}\right),D\right]\cong\frac{1}{N-M}\sum_{\mathrm{n}=M+1}^{N}\mathrm{g}\left(\theta,\sigma_\varepsilon;dose_k,\Delta_{dose}\right),D。 \tag{12}$$

这是在($dose_k$, $dose_k+\Delta dose$]内Logistic 回归曲线可靠度预测值的贝叶斯估计。图3(a) 描绘了样本的判别难度参数b_i取淹没期 (Burn-in Period) $n_0=5\,000$，运行迭代50 000 次的后验均值在不同初始值下的轨迹。由这些轨迹图，我们可以看到样本参数的马尔可夫链最后都在一个值附近作为微小震动，渐进收敛到同一个值，可以说明我们使用的抽样方法是收敛的。图3(b) 描绘了参数θ_1, θ_2, σ_ε的Markov—1D 链计算初始值分别是5.5, −6.3；θ_1, θ_2的建议分布取作相应的正态分布，正态分布函数的方差均为$\sigma^2=0.1^2$；σ_ε的建议分布取作相应的对数正态分布，其方差为$\sigma'^2=1$；Markov —1D 链总长$N=5\times10^3$。

由图4 可知，我们还给出了Logistic 回归模型参数的贝叶斯框架下的50%，90%，95% 和99% 的置信区间 (Credibility Intervals)，其端点由参数的边际后验分布密度函数的4.5% 的分位数和98.5% 的分位数。一般非线性退化轨道模型的性能退化过程，在轨道模型误差项方差未知的情况下，利用贝叶斯估计理论进行可靠性分析与预测。为了在贝叶斯分析框架下考虑未知的误差项方差，采用误差项标准差的无信息先验分布；为了计算预测可靠度函数的后验期望，应用MCMC 方法计算对随机参数及误差项标准差的后验分布进行采样。从这些结果来看，我们提出的基于Logistic 回归建

模和MCMC 方法计算后验描述给定辐照剂量对于丁酸梭菌的作用效果模型对重离子辐射该菌株的真实实验数据的应用，结果是比较合理的。

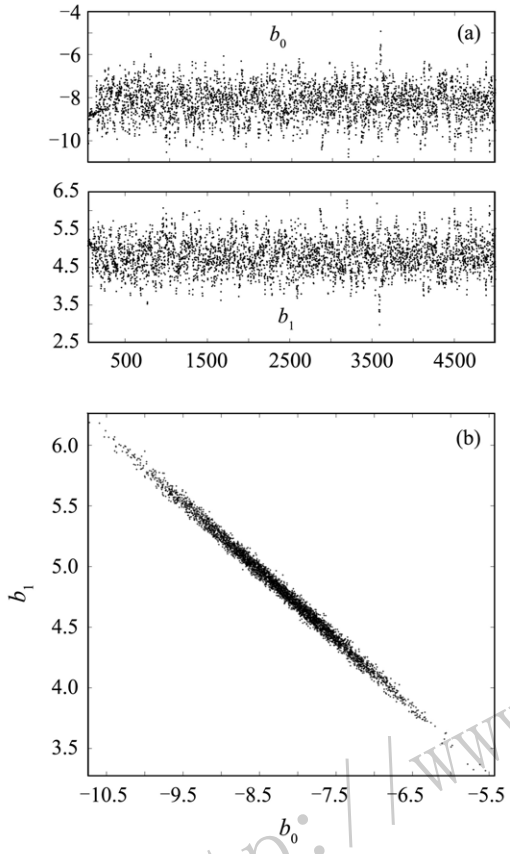

图 3 (a) 表示运行迭代5 000次的后验均值在不同初始值下的轨迹，(b) 表示Markov Chain—1D平稳分布链

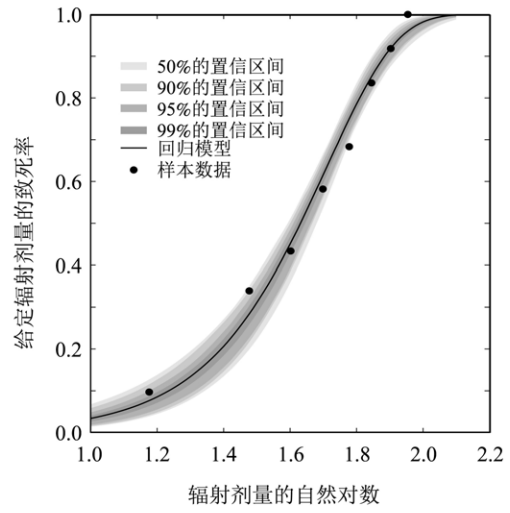

图 4 丁酸梭菌细胞死亡与给定辐射剂量的Logistic回归模型与的贝叶斯框架下的50%，90%，95%和99%的置信区间

特别注意的是，对低剂量的适应性效应，认为这类反应的证据更多地来自对动植物短期效应的观察和培养细胞的研究，然而，有关癌症诱发的大量动物实验的剂量响应关系资料和受低水平照射的人类流行病学调查的有限资料，没有提供可靠证据，证明适应性反应能降低小剂量辐照后诱发这样的晚期效应，即使能降低肿瘤诱发率，但这种现象不总是可以重复发现的。除此之外，Kazama[26]提出对于陆生植物生存率和突变之间的关系——基于有效重离子剂量辐照，其生存率≥ 90% 可诱发突变；Matuo[27]提出对于微生物生存率和突变之间的关系——基于有效重离子剂量辐照，其生存率 10% 可诱发突变。就辐射诱变微生物而言，该模型不仅能够提高重离子束辐照诱变处理的预测精度，又可有效模拟不同种类微生物不同应答变化趋势，避免了大量的辐射实验次数，具有较大科学性与使用性。

4 结论

基于Logistic回归建模和MCMC方法计算后验的理论方法——本文提出的实时可靠性预测方法能有效预测与解释在给定辐射剂量下对于丁酸梭菌株的瞬时杀伤效果，并挖掘出了 Logistic 回归模型参数在贝叶斯框架下的 50%, 90%, 95% 和 99% 的置信区间，提高了重离子束辐照诱变处理菌株的响应精度。该模型方法不仅具有较强的解释能力与很好的通用性，而且可为物理性环境诱变剂 (电离辐射、紫外线、电磁波等) 和化学性环境诱变剂处理微生物制定精度剂量区提供可靠的科学依据与理论指导。

参考文献：

[1] KUTNER M H. Applied Linear Statistical Models[M]. Chicago: Irwin, 1996. **4**: 318.
[2] KUTNER M H, NACHTSHEIM C, NETER J. Applied Linear Regression Models[M]. McGraw-Hill/Irwin, 2004.
[3] PEDUZZI P, CONCATO J, KEMPER E, *et al*. Journal of Clinical Epidemiology, 1996, **49**(12): 1373.
[4] HOSMER J D W, LEMESHOW S. Applied Logistic Regression[M]. John Wiley & Sons,2004.
[5] HARRELL F. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis[M]. Springer, 2015.
[6] MENARD S. Applied Logistic Regression Analysis[M]. Sage, 2002. **106**.
[7] FRIEDMAN J, HASTIE T, TIBSHIRANI R. The Annals of Statistics, 2000, **28**(2): 337.
[8] MOTULSKY H, CHRISTOPOULOS A. Fitting Models to Biological Data Using Linear and Nonlinear Regression: APractical Guide to Curve Fitting[M]. OUP USA, 2004.

[9] POWELL J L. Journal of Econometrics, 1984, **25**(3): 303.
[10] MIN C, ZELLNER A. Journal of Econometrics, 1993, **56**(1): 89.
[11] SIMS C A, ZHA T. International Economic Review, 1998: 949.
[12] GILL J. Bayesian Methods: A Social and Behavioral Sciences Approach[M]. Florida: CRCPress, 2014. **20**.
[13] GEYER C J. Statistical Science, 1992: 473.
[14] GREEN P J. Biometrika, 1995, **82**(4): 711.
[15] GILKS W R. Markov Chain Monte Carlo[M]. John Wiley & Sons, 2005.
[16] HASTINGS W K.Biometrika, 1970, **57**(1): 97.
[17] NEAL R M. Statistics and Computing, 2001, **11**(2): 125.
[18] HAMMERSLEY J. Monte Carlo Methods[M]. Springer Science & Business Media, 2013.
[19] NEAL R M. Journal of Computational and Graphical Statistics, 2000, **9**(2): 249.
[20] METROPOLIS N, ROSENBIUTH A W, ROSENBIUTH M N, *et al*. The Journal of Chemical Physics, 1953, **21**(6): 1087.
[21] GELFAND A E, HILLS S E, RACINE P A, *et al*. Journal of The American Statistical Association, 1990, **85**(412): 972.
[22] QURESHI N, MEAGHER M M, HUANG J, *et al*. Journal of Membrane Science, 2001, **187**(1): 93.
[23] TRAN H T M, CHEIRSILP B, HODGSONod B, *et al*. Biochemical Engineering Journal, 2010, **48**(2): 260.
[24] ZHOU X, XIN Z J, LU X H, *et al*. Bioresource Technology, 2013, **137**: 386.
[25] ZHOU X, LU X H, LI X H, *et al*. Biotechnology for Biofuels, 2014, **7**(1): 1.
[26] KAZAMA Y, SAITO H, YAMAMOTO Y Y, *et al*. Plant Biotechnology, 2008, **25**(1): 113.
[27] MATUO Y, NISHIJIMA S, HASE Y, *et al*. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 2006, **602**(1): 7.

Combining Logistic Regression and Markov Chain Monte-Carlo Describe the Relationship between Exposure to a Given Dose of Radiation and its Effect on *Clostridium tyrobutyricum* Strains

ZHOU Xiang[1,1)], JIANG Tingting[1,2], XU Dan[1], YANG Zhen[1,3], LIANG Jianping[1], WANG Liang[1,2]

(*1. Institute of Modern Physics, Chinese Academy of Sciences, Lanzhou 730000, China;*
2. University of Chinese Academy of Sciences, Beijing 100049, China;
3. Nanjing Agricultural University, Nanjing 210095, China)

Abstract: Using the Markov Chain Monte-Carlo method to estimate the parameters in the Logistic regression model, we constructed a non-periodic irreducible Markov Chain with the posterior distribution of the parameters as stationary distribution, and then used the sample points extracted from the stationary distribution to calculate the Monte-Carlo integral. The above theoretical method can solve the difficult problem of classical logistic regression modeling because of the existence and limitation of the experimental sample data and the multicollinearity. In the classical regression setup with a continuous response, the predicted values can range over all real numbers. Therefore, a different modelling technique is needed. In this work, the results describe in detail a previously unknown lethality trend following $^{12}C^{6+}$ heavy-ion irradiation of *Clostridium tyrobutyricum*. By Markov Chain Monte-Carlo can calculate the model fit for a randomly selected subset of the chain and calculate the predictive envelope of the model. The grey areas in the plot correspond to 50%, 90%, 95%, and 99% posterior regions. More importantly, although this study focused on the use of the method in heavy-ion irradiation of microbial, its results are broadly applicable.

Key words: logistic regression; markov chain monte-carlo; radiation dose; *Clostridium tyrobutyricum*

Received date: 21 Mar. 2016;　**Revised date:** 25 May 2016

Foundation item: CAS Light of West China Program(Ke-Fa-Ren-Zi[2015] No.77); Natural Science Foundation of Gansu Provincial(1506RJZA293)

1) E-mail: syannovich@gmail.com; syannovich@impcas.ac.cn.