建模过程：

- 先只考虑1-mer，即不考虑上下文的情况下的简单模型

  假设n_c是参考序列中核氨酸C出现的总次数，α_{CA}为观测到的C->A的点位数量，以此类推。

  根据多项分布的概率质量函数（PMF），似然函数可以写作：
  $\frac{(n_c)!}{(n_c-n_{CA}-n_{C_T}-n_{CG})!(n_{CA})!(n_{CG})!(n_CT)!}\alpha_{CA}^{n_{CA}}\alpha_{CG}^{n_{CG}}\alpha_{CT}^{n_{CT}}(1-\alpha_{CA}-\alpha_{CG}-\alpha_{CT})^{n_C-n_{CA}-n_{CG}-n_{CT}}$

  这个似然函数表示，在替换总数为 n_C 的情况下，每种核苷酸替换出现的概率。

- 为了找到最优的替换概率估计 $\alpha_{CA},\alpha_{CG},\alpha_{CT}$，我们需要最大化这个似然函数，也就是找到那些能使观测数据 n_{CA},n_{CG},n_{CT} 出现概率最大的参数。通过最大化**对数似然函数**来计算。类似的，我们可以求出$\alpha_{AT},\alpha_{AC},\alpha_{AG},\alpha_{TC},\alpha_{TA},\alpha_{TG},\alpha_{GA},\alpha_{GT},\alpha_{GC}$，最终得到12个概率，根据互补配对原则合并（如A->C与T->G合并），最终得到六个概率。

- 上述的方法可以扩展到3-mer，5-mer甚至7-mer，例如，如果我们要考虑局部序列上下文ACA，那么我们计算 3 聚体序列 ACA 在参考基因组中出现的次数n_{ACA}，$n_{ACA->AAA}$表示中间位置发生 C 到 A 核苷酸变化的位点数。

- 我们针对不同的序列上下文模型使用其共轭先验，即狄利克雷分布，将狄利克雷先验的浓度参数设置为1。使用最大后验 (MAP) 估计来找到所有可能替换的替换概率估计。对于不同窗口大小的模型的比较：

- 1. 采用对数似然比检验（用上述的似然比来比较）：
  $-2ln(L[data|context\ S_1]) + 2ln(L[data|context\ S_2])$
- 2. 贝叶斯因子分析来比较性能：使用之前发现的特定序列上下文模型的替换概率的 MAP 估计值，计算了整体数据的近似后验似然，用这个后验似然之比来比较。
  $\frac{Prob(data|context\ S_2)Prob(context\ S_2)}{Prob(data|context\ S_1)Prob(context\ S_1)}$

- 为了量化7-mer上下文中不同序列组合对替换概率的影响，使用了**前向回归**方法来选择最有用的特征：并为 CpG 环境中的三个可能变化分别创建了一个额外的替代类。初始回归模型：

$$\Pr[X_1 \to X_2|S]=\alpha+\beta_1 p_1^{\mathrm{C}}+\beta_2 p_1^{\mathrm{G}}+\beta_3 p_1^{\mathrm{T}}+\ldots+\beta_n p_7^{\mathrm{T}}+\varepsilon$$

- 添加双向，三向，四向的非线性相互作用

$$\Pr[X_1 \to X_2|S]=\alpha+\beta_1 p_1^{\mathrm{C}}+\beta_2 p_1^{\mathrm{G}}+\beta_3 p_1^{\mathrm{T}}+\ldots+\beta_n p_7^{\mathrm{T}}+ \quad (5)$$

$$\beta_a p_i^w \times p_j^x+\ldots+\beta_b p_i^w \times p_j^x \times p_k^y+\ldots+\beta_c p_i^w \times p_j^x \times p_k^y \times p_l^z+\ldots+\varepsilon$$

  其中β_a后跟的是双向，β_b后跟的三向，β_c后跟的四向相互作用

  对每个复杂程度不断增加的交互级别（首先是双向，然后是三向，最后是四向）进行逐步前向回归。对于每个交互级别，我们通过依次合并交互项（一次一个）来进一步训练模型，并使用ANOVA F检验评估每个项是否改进了模型。

- 我们重复此过程，直到没有其他特征进一步改进模型（所有提出的特征在 ANOVA F检验中P > 0.001 ）。作为我们的最终模型，我们选择了均方误差最小的训练模型，该模型通过每个替换类中的交叉验证计算得出。