# 数学基础

## 连续时间马尔可夫过程

**定义1**我们定义一个随时间连续变化的连续时间马尔可夫过程。与离散时间的马尔可夫链定义相似，我们认为连续时间是连续时间马尔可夫链，如果对于一切和非负整数,, 当 有



换句话说，在连续时间的马尔可夫链是具有与离散时间马尔可夫链相同的性质，即给定**当前状态，未来的状态与过去的状态是条件独立的。**

**定义2** 此外，如果对于任意的和状态，过渡概率满足



则称该过程具有时齐性。这意味着**过渡概率不依赖于过程开始的具体时间，而仅依赖于状态之间的时间间隔**，这是连续时齐马尔可夫过程的关键特征。

* 定义1说的是**马氏性**、**未来只与现在有关，而与过去无关**
* 定义2说的是在某一时刻停留的时间具有**无记忆性**

不妨设某马尔可夫过程在时刻状态停留的时间为转移为状态，该停留时间具有无记忆性，故服从指数分布，不妨设。则平均停留时间为，从状态到状态的转移速率为，可以理解为单位时间内从状态转移到状态的次数，如果在一个很小的时间段,则从状态历经时间转移到状态的概率近似为,省略了一个的一个无穷小量。

**定义3** 设表示马尔可夫过程从状态历经时间转移到状态的转移概率。

**定理1（C-K方程）**对于任意,我们有

证明：



**定义3**设表示从状态历经时间转移到状态的概率的导数

**定义4**设表示从状态转移到**其他不同状态**的转移速率,表示从状态转移到状态的瞬时速率，有,从而

**引理 1(a)** **(b)**()

**定理2(柯尔莫哥洛夫向后方程)** 对于一切状态, 和时间，我们有：



其中，表示转移概率的时间导数，是从状态到状态的转移速率，且 是状态的离开速率。

证明：根据导数定义和引理1以及C-K方程容易证明

**定理2(柯尔莫哥洛夫向前方程)**对于一切状态 , 和时间 ，我们有：



其中， 表示时间时从状态到状态的转移概率，表示从状态到状态的转移速率，且 是状态的总离开速率。

对于任意的状态和，令



则向后方程的矩阵形式为



同样地，向前方程地矩阵形式为



称为**转移速率矩阵**

**该微分方程的解为**,其中(关键在于计算,直接计算或者近似计算都可以，关键在于看矩阵的性质)

## 极大似然估计

极大似然估计的**基本思想**是选择使得观察到的数据出现概率（即似然）最大的参数值。具体来说，给定一组观测数据和一个依赖于某些未知参数的概率模型，MLE的目标是找到这些参数的值，使得这组数据在该模型下的概率（似然）最大（转化为**优化问题**）。极大似然估计是频率派的，认为模型的参数是一个未知的**常量**。

假设有一组观测数据，我们认为它们服从正态分布，但不知道均值和方差。我们的目标是估计和。

似然函数： 

对数似然函数： 

最大化对数似然： 通过求导并解等式和，可以得到 和的MLE估计。

## 贝叶斯估计

贝叶斯估计是统计学中一种基于贝叶斯定理的参数估计方法。与频率主义统计（如极大似然估计）侧重于从数据中直接估计参数不同，贝叶斯估计结合了先验知识（认为参数是**随机变量\向量**）和观测数据来更新对参数的认识，形成后验分布。

贝叶斯定理：



步骤：

**选择先验分布**：基于先前的知识或假设选择一个合适的先验分布。这可以是均匀分布（表示无信息先验）、高斯分布或任何其他反映先验知识的分布。

**构建似然函数**：根据数据生成模型构建似然函数，它表示不同参数下数据出现的概率。

**计算后验分布**：利用贝叶斯定理结合先验分布和似然函数来计算后验分布。这一步通常涉及复杂的计算，特别是当模型参数多或数据结构复杂时。(主要计算就是分母的积分，可以用蒙特卡洛方法计算)因此，贝叶斯估计，将问题转化为**求积分问题**。

**参数估计**：根据后验分布，可以选择后验分布的均值、中位数或众数作为参数的估计值。众数对应于最大后验估计（MAP）。

# 分子进化与系统发育建模

分子进化指的是，每一个DNA序列的每一个位点它的替代服从一个连续时间的马尔可夫过程（最简单的情况，不同位点进化独立同分布、模型参数不随时间变化（即马尔可夫过程的substitution rate is a constant）），该模型有12个参数需要估计，但是如果继续假设该模型是time-reversibility则模型的参数可以减少到9个。

系统发育

DNA序列变化的过程可以被认为是沿着二叉树的分支运行。即当进程到达树的一个节点时，当前状态被复制，两个独立的进程沿着两个子分支重新启动。该过程有参数1.树的拓扑结构（对于一共N个物种来讲，有根二叉树有(2N-3)!!种情况、无根二叉树来说有(2N-5)!!种情况）2.分支的长度（一般表示为给定时间内DNA分子期望替代的次数，即分支越长表示连接的两个node的分子之间的差异越大）

# 例子

给定三条序列

S1：TGG

S2：AGG

S3：AGC

构建进化树

## 极大似然估计

用JC69模型对其进化过程进行建模

**JC69模型的基本假设：**

1. **所有位点独立演化**：序列中的每一个核苷酸位点独立于其他位点演化。
2. **核苷酸变化是均匀和对称的**：从任何一个核苷酸变化到另一个核苷酸的概率是相同的，即任一核苷酸变为其他任一核苷酸的概率是等可能的。
3. **替换率是恒定的**：模型假设替换率在整个序列以及演化过程中是恒定的。

**替代速率矩阵**：



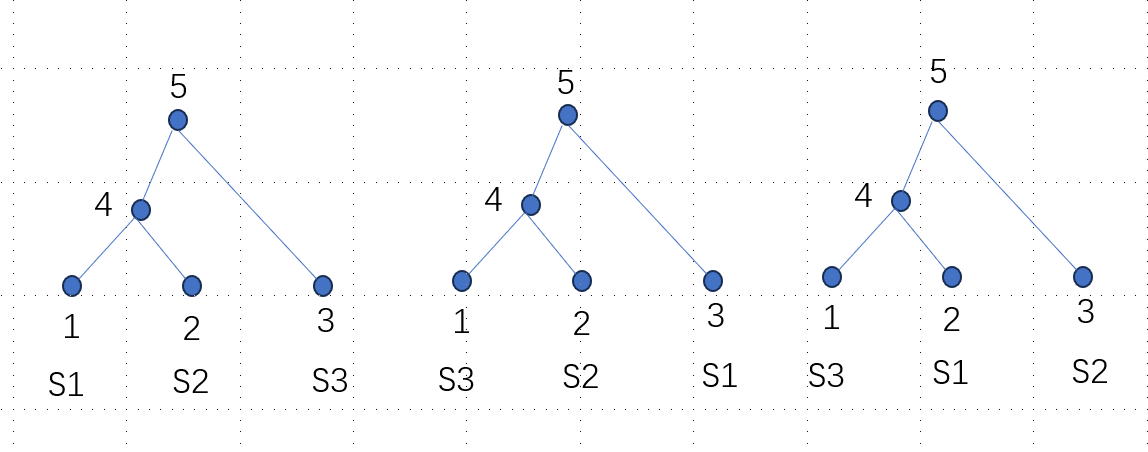
状态转移矩阵：



其中

表示从状态历经时间转移到状态的概率，

有根树的拓扑结构共三种：



**计算似然值的方法：**

似然函数定义为给定模型参数下数据的概率：



在DNA序列数据的情况下，假设序列的每个位点独立演化，似然函数可以分解为每个位点的似然函数的乘积：



其中，每个是序列中第位的数据。对于树中的每一个内部节点，定义条件似然 为在节点处状态为时，从到序列末端所有叶子的观测数据的似然：



其中是位点在所有从节点派生出的序列的集合。

根据Felsenstein的方法，每个节点的条件似然可以递归地计算为：



其中k1和k2是节点 k 的两个子节点，和是从到k1和k2 的分支长度，表示在时间内从状态转移到状态的概率。 整个序列的似然值可以通过根节点的条件似然计算得到，根节点的似然值表示为：



其中是根节点，是根节点处状态的先验概率。递归终止条件



**Tree 1的似然值**

**取****，**,





**只示例计算位点1的似然**

**首先**

根据递归方程：

节点4









**节点5**



**其余类似**