图1、不同核函数计算密度曲线

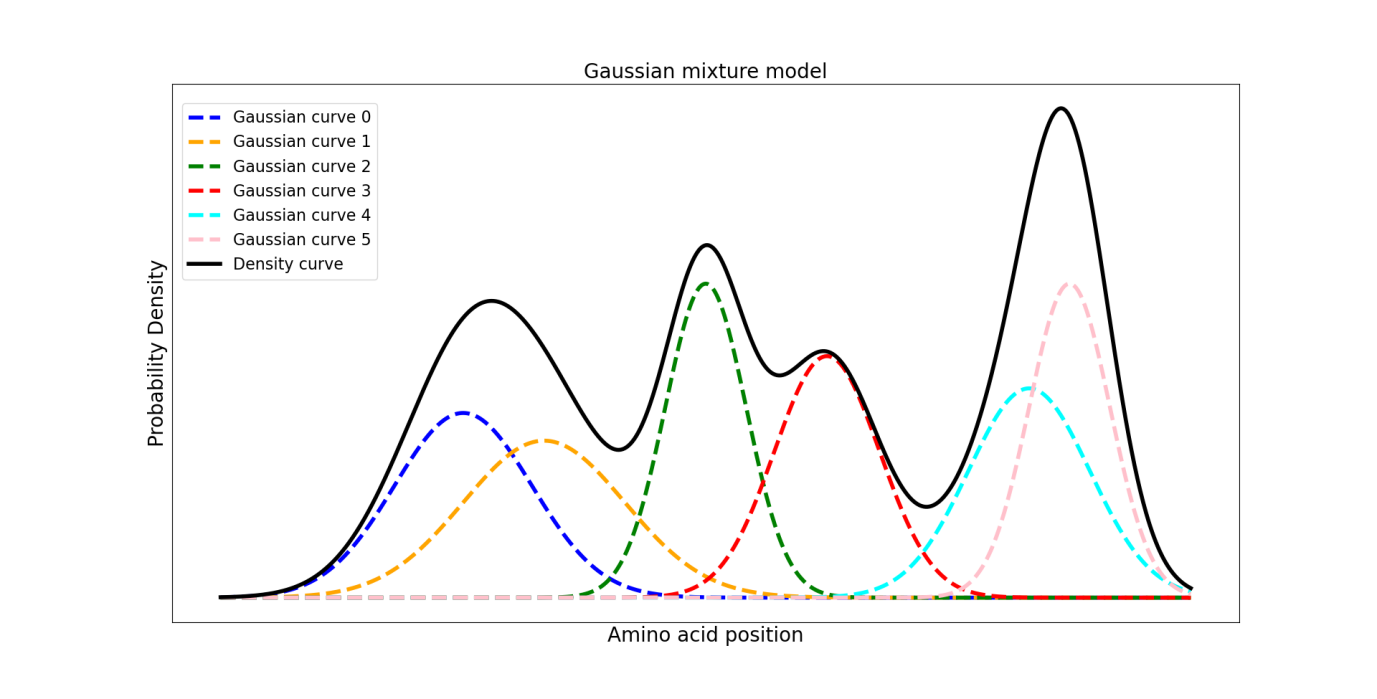


图2、高斯混合模型

**最佳带宽计算步骤**

通过最小化均积分平方误差（*MISE*）获得特定氨基酸序列的最佳固定带宽：

而，其中*AMISE*为渐进的*MISE。AMISE*有：

其中：

最小化*MISE(h)*等价于最小化*AMISE(h)*，求导，令导数为0有：

得：

当核函数确定以后，*h*公式里面得*R、m、*都可以确定下来，*h*便存在解析解。

我们选取的核函数为高斯核函数，所以我们选取的最佳带宽为：

这里 σ是样品的标准差。

**期望值最大算法**

EM算法由3个步骤组成：E步（计算*μ*和*σ*的期望值）、M步（从 E步计算这两个参数对高斯模型的最大似然估计值）和迭代步（重复 E步和M步，直到这两个参数收敛）。这个过程决定了突变簇的边界，通过该边界可以识别一组突变簇。

具体步骤：1、初始化*μ*和*σ*，我们设定核密度曲线的极大值为*(k=1,2,...,K)，K*为极大值个数，每个极大值两侧分别对应了两个极小值，在 *k*之前或之后没有局部最小值的边缘情况下，值*(0,f(0))*和*(L,f(L))*根据情况用作局部最小值。其中 *L* 是基因中氨基酸的长度。然后每个高斯分布的初始参数和由下式给出：

2、计算所有样本对应类别的概率

3、更新和

4、重复第2、3步，直到最大迭代次数或者迭代中的*μ、σ、ω*收敛

完成上面迭代以后，将会得到*J (J<=K)*个*μ、σ*，即可对应*J*个候选热点区[]。

表S1、各条件下基因数量

|  |  |
| --- | --- |
| 类型 | 基因数量 |
| 包含所有热点和冷点的基因 | 899 |
| 仅包含热点的基因 | 343 |
| 仅包含冷点的基因 | 464 |
| 既包含热点又包含冷点的基因 | 41 |
| 含有热点的基因中PLP数量超过PLP+BLB总数80%的基因 | 242 |
| 含有热点的基因中仅包含PLP的基因 | 43 |
| 含有冷点的基因中BLB数量超过PLP+BLB总数80%的基因 | 432 |
| 含有冷点的基因中仅包含BLB的基因 | 277 |
| 所有基因中PLP数量超过PLP+BLB总数80%的基因 | 546 |
| 所有基因中BLB数量超过PLP+BLB总数80%的基因 | 5749 |
| 所有基因中仅包含PLP的基因 | 259 |
| 所有基因中仅包含BLB的基因 | 5290 |