

# 第二问建模思路

2025 年 6 月 1 日

## 摘要

本文针对混合 DNA 样本分析问题，提出了基于贡献矩阵和优化算法的解决方案。方法分为五个步骤：人数分类、构建贡献矩阵、建立优化模型、算法求解和结果验证。重点解决了二人混合样本的比例估计问题。

## 1 人数分类与预处理

根据第一问的建模结果，首先对混合样本进行人数判断和数据处理。本文以二人混合样本为例进行分析，方法可推广至多人情况。

由于第一问的模型建立在等比例的混合样本上，所以应先对第一问的模型加以改进，使得可以更好的判断不同比例的混合样本贡献者人数。

加入比例特征：

1: 峰高偏度 **peak height skewness**

$$\gamma = \frac{\frac{1}{n} \sum_{i=1}^n (h_i - \bar{h})^3}{\sigma^3} \quad (1)$$

其中：

- $h_i$  = 等位基因  $i$  的峰高
- $\bar{h}$  = 该基因座等位基因的平均峰高
- $\sigma$  = 峰高标准差
- $n$  = 该基因座的基因总数

## 2 贡献矩阵构建

考虑二人混合样本，设贡献者分别为 A 和 B。

### 2.1 贡献矩阵定义

贡献矩阵 A 定义为：

$$A = (a_{ij})_{16 \times 26}$$

其中：

- 行数 16 对应数据集中的 16 个基因座

- 列数 26 对应各基因座上可能的基因类型（按数据集顺序排列）
  - 元素  $a_{ij} \in \{0, 1, 2\}$  表示贡献者 A 在第  $i$  个基因座上第  $j$  个基因的贡献个数
- 同理可定义贡献者 B 的贡献矩阵  $B$ 。

## 2.2 混合比例约束

设  $p_1$  为 A 的混合比例， $p_2$  为 B 的混合比例，满足：

$$p_1 + p_2 = 1, \quad p_1, p_2 \geq 0$$

引入基因座混合比例变量：

$p_{1i}$  为 A 在第  $i$  个基因座对应的比例， $p_{2i}$  为 B 在第  $i$  个基因座对应比例

## 3 优化模型建立与求解

### 3.1 比例矩阵定义

引入归一化的比例矩阵：

$$C = (c_{ij})_{16 \times 26}$$

其中  $c_{ij}$  表示第  $i$  个基因座上第  $j$  个基因的归一化峰高值（除以该基因座的总峰高），用于消除基因扩增效率差异。

考虑到直接穷举计算量爆炸，于是采取单基因座穷举结合全局加权平均的方法减少计算量，并且增加精确度。

引入比例矩阵行向量：

$$C_i \quad i \in \{1, 2, \dots, 16\}$$

其中  $C_i$  即为比例矩阵  $C$  的第  $i$  行行向量。

对应的，引入贡献矩阵行向量：

$$A_i \quad i \in \{1, 2, \dots, 16\}$$

其中  $A_i$  即为贡献矩阵  $A$  的第  $i$  行行向量，同样的  $B_i$  表示贡献矩阵  $B$  的第  $i$  行行向量。

### 3.2 优化问题建模

单基因层面目标是最小化预测值与观测值的差异：

$$\min_{p_{1i}, p_{2i}, A_i, B_i} \|p_{1i}A_i + p_{2i}B_i - C_i\|_F$$

其中  $\|\cdot\|_F$  表示矩阵的 Frobenius 范数（各元素平方和的平方根）。 $i \in \{1, 2, \dots, 16\}$

约束条件：

1. 比例约束：  $p_{1i} + p_{2i} = 1$
2. 基因贡献约束：  $\sum_{j=1}^{26} a_{ij} = 2, \forall i = 1, \dots, 16$
3. 元素取值约束：  $a_{ij} \in \{0, 1, 2\}$

### 3.3 求解算法

采用以下方法求解：

- **穷举法**：由于矩阵元素取值有限且问题规模可控，穷举法可行
- **最小二乘法**：用于在给定  $A, B$  时快速求解最优比例  $p_{1i}, p_{2i}$

### 3.4 总比例的求解

根据上述的方法已经求解出各 16 个基因座对应的最优比例，接下来求解整体比例。引入权重项：

$$\omega_i = \sqrt{\sum_{j \in C_i} c_{ij}}$$

该权重项基于峰高对混合比例的强相关性。归一化处理，令

$$\omega'_i = \frac{\omega_i}{\sum_{i=1}^{16} \omega_i}$$

则  $\omega'_i$  即为基因座  $i$  对应的权重系数。由此

$$p_1 = \sum_{i=1}^{16} \omega'_i p_{1i} \quad p_2 = \sum_{i=1}^{16} \omega'_i p_{2i}$$

即为所求混合比例。

## 4 算法可行性分析

- **计算效率**：矩阵维度有限（ $1 \times 26$ ），且每人单基因座基因贡献数固定为 2
- **线性关系**：人数与基因数量呈线性关系，保证算法可扩展性
- **数值稳定性**：归一化处理提高了数值计算的稳定性