

# 第二问建模思路

2025 年 5 月 24 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。

## 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：

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

### 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。

#### 3.2 优化问题建模

目标是最小化预测值与观测值的差异：

$$\min_{p_1, p_2, A, B} \|p_1 A + p_2 B - C\|_F$$

其中  $\|\cdot\|_F$  表示矩阵的 Frobenius 范数（各元素平方和的平方根）。

约束条件：

1. 比例约束：  $p_1 + p_2 = 1$
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  $p_1, p_2$
- 分步优化：先固定  $A, B$  优化比例，再固定比例优化  $A, B$ ，迭代收敛

### 4 算法可行性分析

- 计算效率：矩阵维度有限（ $16 \times 26$ ），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