转录因子预测

1.获取启动子序列

1.1 NCBI Gene数据库 → 输入目标基因（如TMCO1）
1.2 确认转录方向（箭头→表正向，箭头←表反向）
1.3 计算启动子区：正向基因：TSS-2000 bp 至 TSS+100 bp；反向基因：TSS-100 bp 至 TSS+2000 bp
1.4 下载FASTA格式序列（核对基因组版本，如hg38）

2.UCSC添加JASPAR数据库

2.1 UCSC官网 → Track Hubs → 搜索"JASPAR“
2.2 Connect数据库 → 选择匹配基因组版本（如hg38）
2.3 输入启动子坐标 → Search可视化
2.4 过滤设置：Minimum Score > 600

3.预测转录因子

3.1 优先选结合方向一致的TF（如反向基因选“←”，正向基因选“→”）

4.扫描结合位点

4.1 JASPAR官网添加目标TF到购物车 → 点击"Scan"
4.2 粘贴启动子序列 → 设相对阈值90%
4.3分析结果：优先选择结合位点数量多的TF，匹配分数（越高越可靠）

常见问题解析

1. 启动子区域计算错误导致预测失效

问题原因：未正确区分基因转录方向（正向/反向），错误计算上下游坐标（如将反向基因TMCO1的启动子误算为TSS-2000至TSS+100，实际应为TSS-100至TSS+2000）。未考虑基因可变启动子（如多个转录起始位点）。

解决方案：①在NCBI Gene页面确认"Genomic Context"中的转录方向箭头（←表反向）；②反向基因的启动子公式：TSS-100 bp至TSS+2000 bp（以TSS为坐标原点）；③通过UCSC浏览器输入坐标后，检查区域是否包含 "CpG岛"或已知启动子标记（如H3K4me3组蛋白修饰）。

2. JASPAR预测结果与UCSC可视化不一致

问题原因：UCSC中启用了多个JASPAR版本（如JASPAR2020与JASPAR2022），但未统一筛选阈值；数据库版本与基因组版本冲突（如用hg19坐标查询hg38数据库）。

解决方案：①在UCSC的 "Track Controls"中关闭旧版数据库（如隐藏JASPAR2020）；②确保JASPAR分数阈值统一（例：仅显示Score>600的TFBS）；③在JASPAR官网 "Scan"功能中复现UCSC坐标，验证结合位点。

3. Scan功能报错"Invalid sequence format"

问题原因：从NCBI复制的FASTA序列含多余注释行（如>NC_000001.11 Homo sapiens...）；序列包含非标准字符（如空格、数字或小写字母）。

解决方案：①使用纯文本编辑器处理FASTA文件，仅保留">"开头的首行和后续序列；②将序列转为全大写字母，删除所有非ATCG字符；③在JASPAR的Scan页面勾选 "Allow non-standard nucleotides" 选项。

参考文献

[1] Liebold J. Transcription factor prediction using protein 3D secondary structures. Bioinformatics, 2024, 41(1): btae762.