Prediction of Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs

摘要：

脱靶效应可导致次优基因编辑结果，是其发展的瓶颈。

使用基于两个相互关联的**机器学习模型**的方法来预测脱靶效应---叫做Elevation。

对独立的guide-target 对进行评分，然后同时将他们合并为一个唯一整体总结指导分数。还提出了一张评估方法用于：平衡活动和非活动guide之间的误差

背景：

减少脱靶影响最好的方式是：知道他们什么时候，在哪发生，并在平衡on-target效率的情况下设计一个指南来避免脱靶。

本文提出了基于机器学习的方法：

基于机器学习的预测建模可以利用少量的数据来了解导致脱靶效应的gRNA-target序列对的统计规律，以及它们对细胞的总体影响。这种建模使得能够廉价和快速地在全基因组水平上筛选非试验前的gRNA的脱靶效应。

本文方法：

对于脱靶预测建模有两种主要的用例： 1，了解给定的脱靶区域可能对特定的gRNA来说有多活跃，这个活跃性称之为gRNA-target评分。（对于关注基因组的特定区域来说是很有用的） 2，获得给定的gRNA的所有脱靶活性的总体得分，以获得基因的潜在gRNA排序。

可以将脱靶预测模型问题分解成三个主要的任务：

- 搜索并过滤全基因组以获得一个gRNA的潜在靶点。 （例如，基因组中与gRNA相匹配的区域中，可能有N个*目标位点*的核苷酸误匹配，这些位点将会在第二步以后才会被视为脱靶活性，使用机器学习去区分有活性的和无活性的targets。在这一步只会创建一个潜在活性位点的简短列表。）
- 对每一个潜在活动目标进行评分，给gRNA-target对分配一个数值，来表示一个gRNA-target对预计有多少脱靶效应。
- 对（2）中的分数进行合并，得到一个单独的脱靶可能性，用于评估gRNA。

在第一步的搜索和过滤任务中，可以使用数值方法：Cas-OFFinder, CRISPOR, CHOP-CHOP, e-CRISPR， CRISPR-DO, CROP-IT and COSMID。（所算法使用的搜索算法不同，以及有着不同的*搜索完整性*。）

搜索完整性：取决于诸如最大不匹配数量，允许的原型间隔符相邻基序（PAM）和所使用的搜索算法。

本文中，在第一步中使用了自己的系统作为搜索和过滤操作----*Elevation-search*。

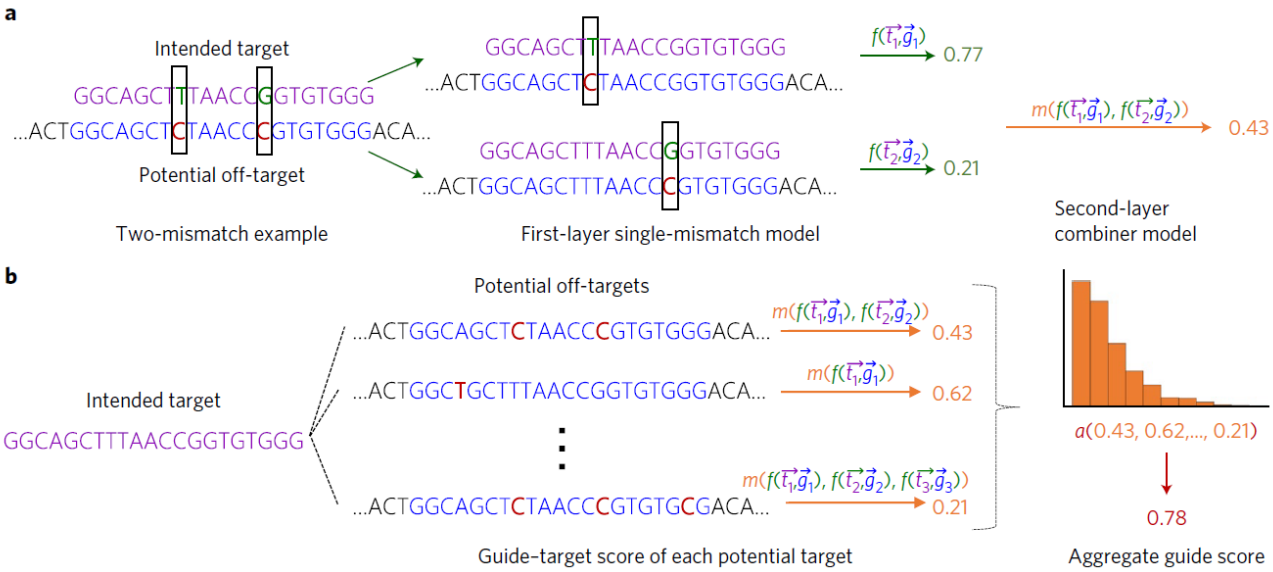
第二步和第三步使用所提出的---脱靶的端对端建模方法为**Elevation**：

- 对于gRNA-target评分：开发出一个**双层回归模型**（**Elevation-score**）：
  - 第一层学习预测单个错配（target与预期的target之间，因此也包括替代的PAM）脱靶活性。
  - 第二层模型学习如何将来自具有多个错配的gRNA-target对的单错配模型的预测组合成单个gRNA-target得分—-**‘combiner’模型**。
- 为了汇总guide的单个评分：
  - 首先将gRNA-target评分模型应用于潜在的target列表（通过Elevation搜索列出）。
  - 然后再使用Elevation-汇总模型来对单个的评分进行汇总。
    *Elevation模型：考虑每一个潜在的target是否位于基因中，并允许这些特征和其他特征通过非线性建模方法（**boosted 回归树**）相互作用。

下图为Elevation脱靶预测模型的框架图：

- 第一层*a*：
  - 首先gRNA-target对被分解为两个单个误匹配pseudo-pairs（{}t_1, g_1},{t_2, g_2}），每一对可以通过第一层（单个误匹配）模型来获得评分f。
  - 然后这些单个的评分通过第二层模型进行组合，生成一个解释所有误匹配的单个gRNA-target评分。

- 第二层*b*：
  - 计算gRNA-target评分的输入分布的统计作为特征，并通过模型运行产生一个gRNA的聚合评分。



特征选择：

对于第一层（单个误匹配）特征选择：

- 误匹配的位置.
- 误匹配的nucleotide（核苷酸）一致性(the nucleotide identity）.
- 单个特征中无匹配的联合位置和身份.
- 突变是否是一个转换或者颠换。

第二层（多个误匹配组合器）模型：

- 特征重要性显示误匹配的总数，以及第一层单个误匹配预测的总和在驱动了这个模型。

聚合的最后任务：获得对于一个gRNA所给定的所有单个gRNA-target评分的单个脱靶总结评分。该任务的解决方案对于gRNA的设计非常有用，在于用户想要扫描大量的gRNA的总体活性。

所用到的机器学习算法：

- 双层回归模型
- boosted回归树