名词:

**RNA methylation** : RNA甲基化 真核生物丰富和调节遗传信息的一种保守的转录后机制

**Three Kingdoms of Life** : bacteria, archaea and eukaryotes 细菌、古细菌和真核生物

**Methyl-groups**: 一个由一个碳原子和三个氢原子组成的小分子。甲基从蛋白质或核酸中被添加或去除，并可能改变这些分子在体内的作用方式。

**Up-regulation**:基因表达上调 上调就是基因转录成mRNA时受到正向调控,促进表达

摘要:

用五个分类器模型，在多个训练集上使用交叉验证，验证集达到88%，测试集达到91%

提出6个分子亚网络 将预测数据和已知RNA甲基化基因链接起来 能够发挥作用。

**研究例证了通过机器学习的方法在大型生物数据集上预测基因功能**

处理：

1. 确定92种与RNA甲基化有关的蛋白质(已知的甲基转移酶、先前被注释为假定的转移酶)
2. 为了平衡样本的问题，将2W多个负样本分割大小等同(74)的290个训练集
3. 取Harmonizome Dataset中的Gene-Attribute Matrix 十多个数据集组合成一个文件，筛选低方差特征，降低到1505个特征，事实上去掉了HubProtein KEGG Reactome 三个数据集的所有特征通过排除GO特征，组装了第二个降维特征集
4. GB模型预测的前100个基因于已知的甲基化相关基因做层次聚类分析
5. 在STRING数据库中查询已知RNA甲基化基因和人类基因组其他基因的独立蛋白质-蛋白质相互作用(PPI)信息。 建立一个基于相互作用的PPI网络。（这个网络怎么建立的没说）通过PageRank算法对网络进行评分，将评分作为输入进行基因集富集分析Gene Set Enrichment Analysis (GSEA) ，用这个实验证明预测的基因和甲基化密切相关，鲁文群落检测法，确定六个社区
6. 为了处理GO注释特征表达信息丰富问题，去掉所有GO特征作为一个降维数据集进行训练，实验结果表明没什么太大区别
7. 详细说明六个社区意义

其他:

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Precision = TP/(TP+FP)

Recall = TP/(TP+FN)

F1 = (2\*Precision\*Recall) / (Precision + Recall)

梯度提升: 每次训练一个弱分类器拟合上一步的误差，将训练出的多个分类器加起来得到一个更强的分类器。为了控制过拟合加入学习率。

集成学习： 多个模型解决一个问题

逻辑回归：找一个线性函数拟合sigmoid函数,本文优化器选择牛顿法