# Effective small interfering RNA design based on convolutional neural network

# 摘要：

在功能基因组学中，小干扰RNA（siRNA）可用于敲低基因表达。通常，目标基因有许多潜在的siRNA，但它们的基因沉默效率通常是不同的。因此，对于成功的RNA干扰（RNAi），选择最有效的siRNA是关键的一步。尽管已经开发了各种计算算法，但功效预测的准确性并不令人满意。为了探索不同基序对基因沉默效应的影响，进一步提高预测精度，我们利用深度学习算法 - 卷积神经网络（CNN）开发了一个新的强大预测器。对比结果表明，我们模型的Pearson相关系数（PCC）为0.717，比Biopredsi，i-Score，ThermoComposition21和DSIR高13.81％，16.78％和5.91％。此外，我们模型的ROC曲线下面积（AUC）为0.894，比这四种算法高10.10％，12.59％和7.07％。结果显示，我们的模型对于预测siRNA沉默功效是稳定且有效的。

**关键词:siRNA; 深入学习; 卷积神经网络;RNA干扰**

# 简介

RNA干扰（RNAi）是小干扰RNA（siRNA）诱导序列特异性转录后基因沉默的过程[1-3]。 RNAi在许多真核系统中被发现，包括植物，真菌，无脊椎动物和哺乳动物[4]。 在哺乳动物细胞中，双链RNA（dsRNA）被加工成短21-23核苷酸（nt）的siRNA并诱导即时靶基因敲低[3]。 迄今为止，siRNA已成为研究基因功能的重要工具[5-7]，为治疗流感病毒[8]，HIV病毒[9]甚至癌症[10]提供了新技术。 与其他基因工具相比，基于siRNA的基因沉默因其简单和低成本而变得越来越流行。

在RNAi的过程中，siRNA设计是最关键的步骤之一，并且在这一领域正在做出许多努力。开发了许多基于机器学习算法的功效预测方法。 Huesken [11]基于人工神经网络开发了一种基于人工神经网络的siRNA功效预测工具'Biopredsi'，并通过高通量分析技术构建了包含2431个siRNA的主要siRNA数据集，开创了siRNA功效预测领域的新纪元。许多siRNA功效预测算法使用Huesken的数据集建立[12]。 ThermoComposition-21 [13]也是一种人工神经网络算法，它结合了热力学和成分特征。为了设计准确且易于解释的功效预测模型，Jean-Philippe Vert [14]提出了一种简单的线性方法，结合了两组简单的siRNA序列特征，包括siRNA序列中每个位置上存在的核苷酸和siRNA的全局内容简短的图案。简单线性模型的功效预测准确度与Biopredsi神经网络一样精确。 i-score [15]也是基于线性回归模型的简单siRNA功效预测算法。该算法专门由核苷酸偏好分数组成，并且可以预测与Biopredsi，ThermoComposition-21和DSIR相似程度的活性siRNA。

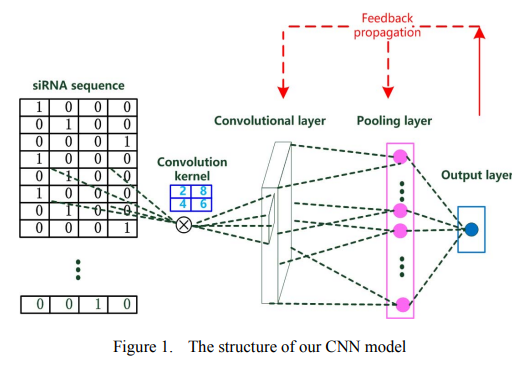
**尽管付出了相当大的努力，但功效预测的准确性并不令人满意。 大多数机器学习算法高度依赖于siRNA特征，包括核苷酸偏好，核苷酸组成，热力学稳定性，二级结构等。 预测能力受到这些特征的限制。**

我们开发了一种基于卷积神经网络（CNN）的新型强大预测器[16,17]。 与以前的功效预测方法不同，我们的模型利用CNN算法自动学习主题编码特征。 在我们CNN模型的卷积层中，卷积核设计为主题（概念）检测器，以数据驱动的方法自动学习siRNA多模主题的潜在特征模式，这种模式更抽象，更贴近本质，更有利于分类。 与当前的siRNA功效预测方法相比，我们的模型在预测准确性方面表现最佳。

# 方法

## CNN模型

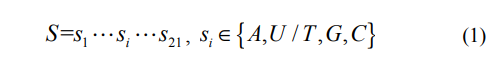
在本文中，我们开发了基于卷积神经网络的siRNA沉默功效预测模型。 结合样本大小和计算复杂度，我们定义了三个重要层，包括卷积层，池化层和输出层。 我们模型的结构如图1所示。



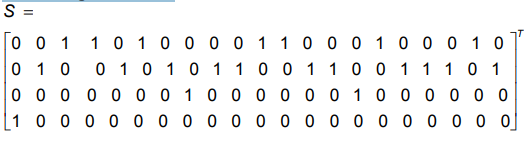
**在卷积层中，设计了多个大小不同的卷积核作为模体检测器，寻找对siRNA沉默效率有重要意义的基序。卷积核可以被认为是从siRNA多模基序的位置和组成中提取的权重。卷积结果是相应的主题编码特征。与经验定义的现有序列编码规则不同，主题编码特征由大型数据集进行训练，具有更多的可用性，指导质量和信息。在池化层中，卷积层的输出被减少。并且选择所有图案的最具代表性的特征模式作为特征表示。最后，逻辑回归函数实现了siRNA沉默功效预测。由于太多的图层产生大量的权重，根据现有siRNA数据集的规模，我们的CNN模型仅包含一个卷积层和一个合并层以检测siRNA多模基序的潜在特征模式，以及一个输出层实现siRNA沉默效能预测。**

## CNN模型编码方法

siRNA序列由21个碱基组成，可以表示为：



在本文中，为了执行卷积操作，将siRNA序列编码成二维矩阵作为我们的CNN模型的输入。 四个碱基A，U / T，G和C以四维二进制形式表示，例如A =，U / T =，G =和C =。 然后将siRNA序列转化到21h4基质中。 例如，S = CUAAUAUGUUAAUUGAUUUat，可以表示为：

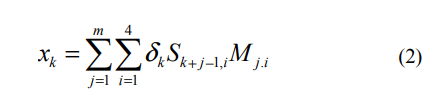


## 主题探测器设计

**由于卷积操作可以检测局部特征，我们设计了各种卷积核探索siRNA多模基序中包含的潜在特征模式。 在CNN模型的训练阶段，利用反向传播算法对大规模训练样本进行卷积核的权重校正，从而保证提取有效的特征模式。**

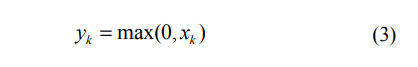
根据siRNA序列编码方法，将siRNA序列转化为21X4矩阵，每个碱基由四维二进制码表示。 由于卷积核的大小不能大于输入矩阵的大小，我们使用大小为mx4的卷积核作为主题检测器。 m表示基序检测器的长度，其值在2至20的范围内。各种大小的卷积核用于检测不同长度的多模基序对siRNA沉默功效预测的贡献。 设计的卷积核可以检测多模基元的局部特征和全序列的全局特征。

进行mx4卷积核M与siRNA序列S之间的卷积运算，并且可以计算卷积层的神经元Xk如下：



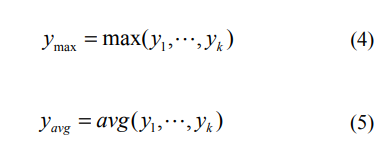
1≤k≤ 22-m，δk是权重修正的学习率。 卷积结果的大小为（22-m）x1，它代表每种多模主题的特征模式。

然后用ReLU函数作为激活函数来增加卷积层中的非线性因子。 比较实验显示ReLU功能表现最佳。 输出yk如下所示：



池化层用于选择最具代表性的卷积结果并去除不相关的信息。 为了保持最显着的局部特征表示和卷积结果的全部信息，本文实现了池化阶段的两种选择：最大池化和平均池化。

在合并操作之后，卷积结果被减少为2维向量y=(ymax,yavg)



因为需要不同的卷积核来检测不同大小的多模基序对siRNA沉默效率的贡献，所以我们的卷积神经网络输出用于回归预测的二维矢量，其中d是卷积核的数目。

## siRNA沉默效率预测的逻辑回归