# 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。

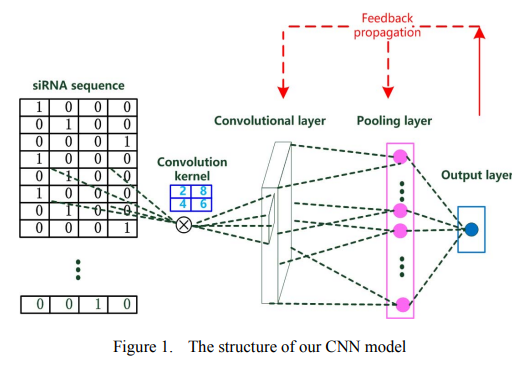
**尽管付出了相当大的努力，但功效预测的准确性并不令人满意。 大多数机器学习算法高度依赖于siRNA特征，包括核苷酸偏好，核苷酸组成，热力学稳定性，二级结构等。 预测能力受到这些特征的限制。**

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

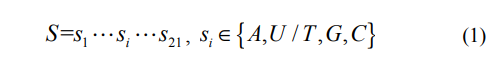
在本文中，我们开发了基于卷积神经网络的siRNA沉默功效预测模型。 结合样本大小和计算复杂度，我们定义了三个重要层，包括卷积层，池化层和输出层。 我们模型的结构如图1所示。



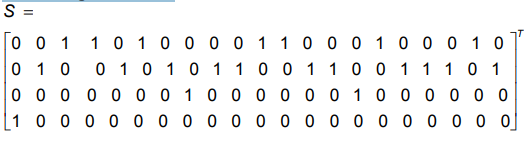
**在卷积层中，设计了多个大小不同的卷积核作为模体检测器，寻找对siRNA沉默效率有重要意义的基序。卷积核可以被认为是从siRNA多模基序的位置和组成中提取的权重。卷积结果是相应的主题编码特征。与经验定义的现有序列编码规则不同，主题编码特征由大型数据集进行训练，具有更多的可用性，指导质量和信息。在池化层中，卷积层的输出被减少。并且选择所有图案的最具代表性的特征模式作为特征表示。最后，逻辑回归函数实现了siRNA沉默功效预测。由于太多的图层产生大量的权重，根据现有siRNA数据集的规模，我们的CNN模型仅包含一个卷积层和一个合并层以检测siRNA多模基序的潜在特征模式，以及一个输出层实现siRNA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

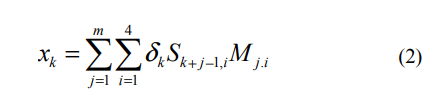


## 主题探测器设计

**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。**

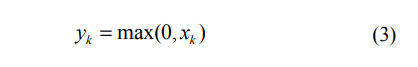
根据siRNA序列编码方法，将siRNA序列转化为21X4矩阵，每个碱基由四维二进制码表示。 由于卷积核的大小不能大于输入矩阵的大小，我们使用大小为mx4的卷积核作为主题检测器。 m表示基序检测器的长度，其值在2至20的范围内。各种大小的卷积核用于检测不同长度的多模基序对siRNA沉默功效预测的贡献。 设计的卷积核可以检测多模基元的局部特征和全序列的全局特征。

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：



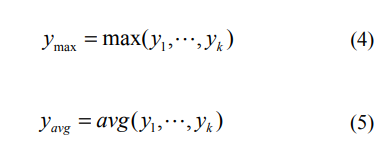
1≤k≤ 22-m，δk是权重修正的学习率。 卷积结果的大小为（22-m）x1，它代表每种多模主题的特征模式。

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)

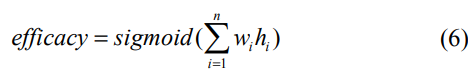


因为需要不同的卷积核来检测不同大小的多模基序对siRNA沉默效率的贡献，所以我们的卷积神经网络输出用于回归预测的二维矢量，其中d是卷积核的数目。

## siRNA沉默效率预测的逻辑回归

激活功能在池化层中实现。 由于定量siRNA沉默效率预测是一个回归问题，输出层中只有一个神经元。 我们开发了一个逻辑回归模型来实现预测过程。 逻辑回归模型是在池化层的神经元之间执行线性加权并连接权重wi，然后通过sigmoid函数映射到输出值。

通过sigmoid函数将[-∞+∞，]范围内的输入值映射到[0,1]范围内的值，这与siRNA沉默效率的范围一致。 而sigmoid函数的中心区域具有较高的增益，两个边界区域的增益较低，适合消除特征向量中由奇点引起的误差。 输出层计算如下：



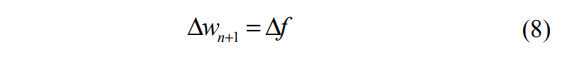
hi是池化层的输出，wi是连接权重。

## 基于CNN算法的siRNA沉默效率预测模型的训练过程

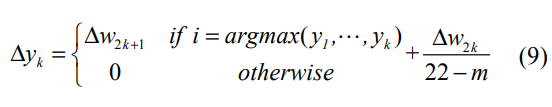
卷积神经网络的梯度由反向传播算法计算。 在本节中，我们根据输出层损失函数的近似最小原理估计误差值。 然后迭代校正卷积层的卷积核和输出层的权重。

当前输出值和标签之间的误差是 Δf。 由于输出层的激活函数为S型函数，因此池化层中连接权重wi的梯度向量Δω如下所示：

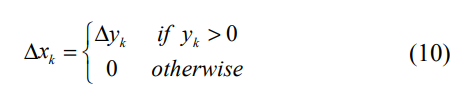




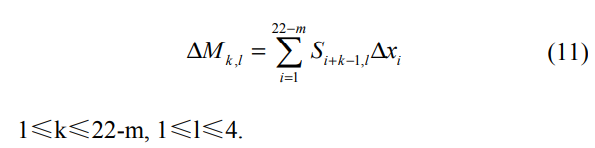
我们实施了池化阶段的最大池化和平均池化。 由于没有模型参数，所以反馈数量计算如下：



卷积层是前向传播的第一层和后向传播的最后一层。 根据来自池化层的反馈量校正卷积核的权重。 由于卷积层的激活函数为RELU函数，因此计算卷积层的神经元Xk的校正如下：



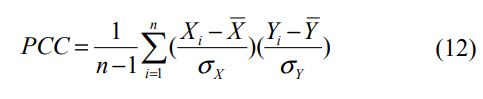
因此，卷积核的权重校正计算如下：



## 评估预测系统

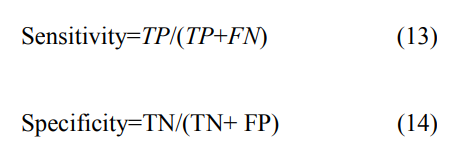
本文采用Pearson相关系数（PCC）和受试者工作特征（ROC）曲线来评估预测模型的性能。

PCC描述了预测模型观察到的和预测的siRNA活性之间的相关性，可以将其定义为：



其中n是样本大小。 Xi和分别是观测值和平均值。

ROC曲线被广泛应用于比较生物信息学中不同算法的效率，并通过绘制灵敏度与1-特异性相关来生成。 他们的定义如下：



其中TN是真正的负标签; FN是假阴性的数量; TP是真正的正正标签，FP是假阳性的数量。 ROC曲线下面积（AUC）可用于测量预测算法的整体性能。

# 结果

## 数据集

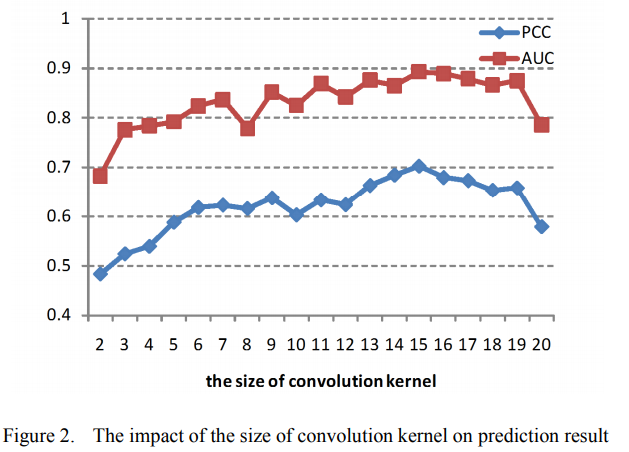
用于比较实验的数据集包括4067个来自Huesken数据集[11]，Reynolds数据集[18]，Vickers数据集[19]，Haborth数据集[20]，Takayuki数据集[21]，Ui-Tei数据集[22] ]和siRNAdb数据集[23]。 使用10倍交叉验证来选择最佳超参数。 在每个实验中，随机抽取407个siRNA样品作为测试集，剩余的3660个siRNA样品作为训练集以开发训练模型。

## 超参数设置

卷积网络的超参数包括卷积核的大小，激活函数，学习速率等。 这些超参数决定了网络结构并直接影响模型的鲁棒性。 由于没有先验知识来指导超参数的设置，因此设计了所有超实验的对比实验，以观察这些超参数对预测结果的影响，并总结超参数设置的规则。

## 卷积核大小对预测结果的影响

在本节中，我们选择不同大小的卷积核来学习不同长度的多模基元的特征表示。 由于siRNA序列长度为21nt，因此有19个多模基序，长度为2~20个。因此，我们采用了19个卷积核学习多模基元的特征。 卷积核的大小是mx4，而2<=m<=20。 在我们的实验中，根据m的值构造了19个卷积神经网络，并且在相应的卷积神经网络中有2-m+1个卷积核作为主题检测器以学习多模的特征模式。 实验结果如图2所示。

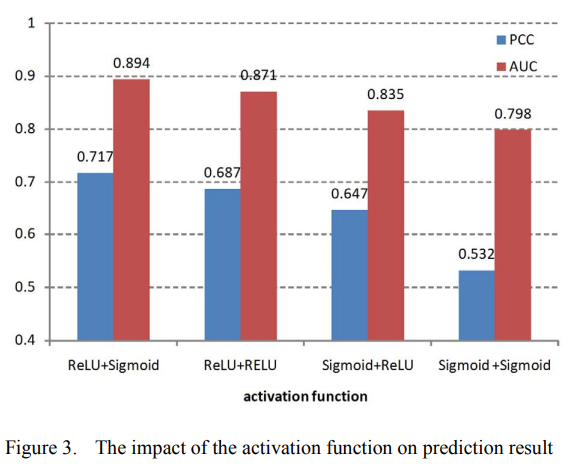


如图2所示，卷积核的大小影响了预测结果。随着m的增加，由相应的卷积神经网络预测的PCC和AUC变大。当m = 15时，PCC和AUC达到最高值。结果表明，我们提出的基序检测器可以从siRNA序列中获得有效的特征模式。卷积核的较小尺寸只能检测与低模主题有关的信息，而忽略高模主题的贡献和全局序列特征。大卷积核可以覆盖更多的多模基元信息，提高预测结果。但如果卷积核过大，卷积神经网络将更注重检测高模主题的特征表示，而忽略具有局部重要性的单核苷酸和低模主题的贡献，从而导致预测结果降低。因此，合理的卷积核大小直接影响有效主题特征学习。获得0.6的PCC的卷积核将作为主题检测器，保证学习特征模式具有足够的判别能力。

## 激活函数对预测结果的影响

在本节中，我们比较了不同激活函数对预测结果的影响。 由于激活函数影响输入信号映射到特征空间的合理性和区分能力，本节将考察不同激活函数的预测效果。 卷积层和池化层有两个激活函数。 卷积层的常用激活函数包括tanh函数，sigmoid函数和RELU函数。 由于tanh函数更适合于特征值存在较大差异的情况，因此不利于检测图案局部微小特征。 本文中，sigmoid函数和RELU函数适用于卷积层，sigmoid函数和RELU函数适用于输出层。 图3给出了卷积层和池化层不同激活函数的预测结果。

图3显示激活函数的选择对预测效果有相当大的影响。 最好的组合是卷积层的激活函数是ReLU函数，输出层的激活函数是S形函数。 ReLU将稀疏添加到从卷积核提取的特征表示中，这可以改进包含在非零神经元中的信息量。 合并层去除了卷积结果中包含零值的神经元，这可以保证基元检测器获得多模基元最具判别性的特征模式。 因此，对于卷积层，ReLU比S形更适合。 在输出层中，sigmoid可以总结每个特征分量的贡献并输出范围从0到1的值。而且，ReLU的输出值范围为0到+∞，这不适合作为输出的活动函数层。



## 学习率对预测结果的影响

学习率是控制校正权重的系数。 如果网络权重可以收敛到最优网络参数，则学习速率的价值会受到影响。 对于不同的输入信号，自适应学习率需要通过实验进行调整。 如果学习速率过大，网络将错过最佳网络参数并陷入局部极值。 然而，学习率太低会导致收敛速度慢，对纠错不敏感。 在这个实验中，我们选择不同的学习率进行比较实验。 如果迭代时间超过1000或者错误小于0.001，则培训可以终止。 根据我们收集的有关卷积神经网络的文献，我们选取0.5,0.1,0.01和0.001作为候选集，分析学习率对预测结果的影响。 实验结果如图4所示

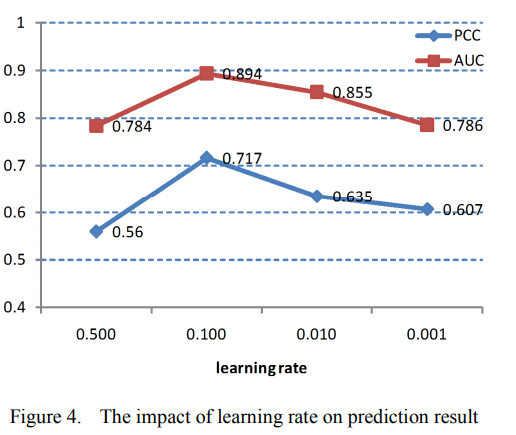


图4显示当学习率为0.1时，预测结果最好。 当学习率为0.5时，预测结果最低，表明网络训练错过了最佳权重并陷入局部极值。 当学习率为0.01和0.001时，PCC和AUC的值相对较低。 该结果表明，当迭代次数为1000时，网络收敛速度较慢，不能获得最优权值。 考虑训练时间和预测结果，卷积神经网络模型的学习率设为0.1。

## 与其他算法相比

根据以上结果，在我们的CNN模型中有14个大小为6x4到19x4的卷积核来检测主题特征模式。 卷积层的激活函数为ReLU，输出层激活函数为S型，学习率为0.1。 在本文中，我们将我们的模型与其他机器学习方法进行了比较，包括siRNApred [24]，Biopredsi [11]和DSIR [14]。 图5显示了五种方法的PCC和AUC值。

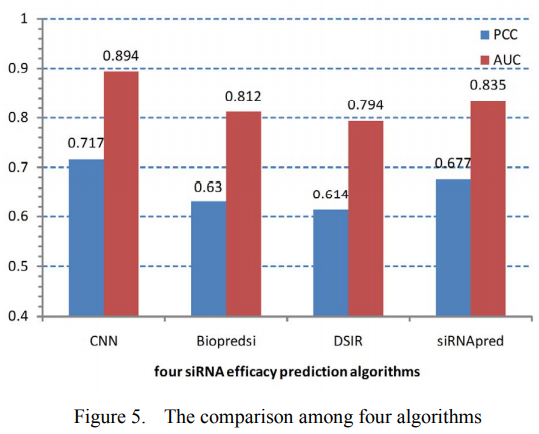


图5显示我们的方法PCC为0.717，高于Biopredsi，DSIR和siRNA分别为13.81％，16.78％和5.91％。 结果表明CNN模型能够有效地学习多模基元的潜在特征模式。 由于映射权重是通过训练数据完全获得的，因此可以更全面地反映siRNA序列特征并获得更好的预测结果。此外，本方法的AUC为0.894，高于Biopredsi，DSIR和siRNA分别为10.10％ 12.59％和7.07％。 结果表明，我们的方法在siRNA沉默效率预测中具有更大的优势，并且比其他机器学习方法更加稳定和有效。

# 总结

siRNA已成为研究基因功能和新的鉴定药物靶点的分子工具。 已经做出许多努力来设计活性siRNA。 因此，开发一种预测siRNA沉默效率的有效方法非常重要。 在这项研究中，我们开发了一种基于卷积神经网络的siRNA功效预测方法。 与现有方法Biopredsi，i-Score，ThermoComposition21和DSIR相比，我们的方法显示出更好的能力。 结果表明，我们的模型可以探索siRNA多模基序在功效预测中的贡献，并利用序列局部特征中包含的有价值信息充分提取特征模式。 鉴于此，数据驱动的特征学习模式比取决于专家知识的学习模式具有更好的性能。