(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111798921 B
(45)授权公告日 2022.08.05

(21)申请号 202010571759.0

(22)申请日 2020.06.22

(65)同一申请的已公布的文献号
申请公布号 CN 111798921 A

(43)申请公布日 2020.10.20

(73)专利权人 武汉大学
地址 430072 湖北省武汉市武昌区珞珈山武汉大学

(72)发明人 杜博 刘子翼 罗甫林

(74)专利代理机构 武汉科皓知识产权代理事务所(特殊普通合伙) 42222
专利代理师 罗飞

(51)Int.Cl.
G16B 5/20(2019.01)
G16B 15/20(2019.01)
G16B 20/30(2019.01)
G16B 40/00(2019.01)
G06N 3/04(2006.01)

(56)对比文件
CN 111192631 A,2020.05.22
CN 111179217 A,2020.05.19

审查员 乔帅

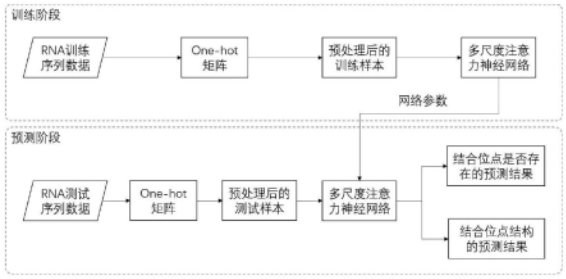
权利要求书2页 说明书11页 附图1页

(54)发明名称

一种基于多尺度注意力卷积神经网络的RNA结合蛋白预测方法及装置

(57)摘要

本发明公开了一种基于多尺度注意力卷积神经网络的RNA结合蛋白预测方法,包括训练阶段和预测阶段。训练阶段包括RNA数据的预处理,RNA数据的编码,构建神经网络和网络参数训练。通过将RNA的数学抽象的统计模式转化成矩阵的形式,输入到预先设计好的基于注意力机制的多尺度卷积神经网络,通过使设计的特别交叉熵损失函数最小,使用Adam优化方法训练神经网络中的参数。在预测阶段,以四个碱基为基本单元的RNA序列数据被输入到网络中,神经网络最后一层输出RNA数据中是否有结合蛋白对应的结合位点的概率大小,从而获得对RNA序列类别的预测结果。本发明可以提高预测精度。



CN 111798921 B

1.一种基于多尺度注意力卷积神经网络的RNA结合蛋白预测方法,其特征在于,包括:

S1:获取RNA数据并进行预处理;

S2:对预处理后的RNA数据进行编码,构建网络训练样本;

S3:构建多尺度注意力卷积神经网络,其中,多尺度注意力卷积神经网络包括多个分支,每个分支设置不同大小的卷积核,分别用以学习在RNA数据中的不同尺度的特征,并引入通道注意力机制学习不同通道在分类时的重要性,在进行RNA结合位点识别时,不同的通道的卷积核对应不同的结合位点结构;

S4:将网络训练样本输入构建的多尺度注意力卷积神经网络中,并采用Adam优化方法训练多尺度注意力卷积神经网络中的参数,得到训练好的多尺度注意力卷积神经网络;

S5:将待预测的RNA数据进行预处理和编码后输入训练好的多尺度注意力卷积神经网络中,得到预测结果,其中,预测结果包括RNA数据是否有结合蛋白对应的结合位点;

其中,S3中构建的多尺度注意力卷积神经网络包括四个分支,第一个分支包括卷积、池化、相乘、卷积、池化和相乘,第二个分支、第三个分支以及第四个分支均包括卷积、池化、卷积和池化,第一个分支的第一个相乘为第一次卷积池化后的输出结果与各通道注意力权重相乘,第二相乘为第二卷积池化后的输出结果与各通道注意力权重相乘;每个分支提取出的不同尺度的特征相加后,通过一个全连接层,得到最终预测结果。

2.如权利要求1所述的预测方法,其特征在于,S1具体包括:将获取的不同长度的RNA数据填补至相同的长度。

3.如权利要求1所述的预测方法,其特征在于,S2具体包括:

将预处理后的RNA数据采用One-hot矩阵表示,构成网络训练样本。

4.如权利要求1所述的预测方法,其特征在于,构建的网络训练样本包括正训练样本和负训练样本,蛋白质对应的正训练样本为包含有该蛋白质结合位点的RNA数据,负训练样本为无该蛋白质结合位点的RNA数据,训练过程中,正训练样本标签赋值为1,负训练样本标签赋值为0。

5.如权利要求1所述的预测方法,其特征在于,通道注意力权重的计算方式为:

$$Z_k = \frac{1}{W}\sum_{i=1}^{W} X_{i,k}$$

$$outputs = softmax(W_2 sigmoid(W_1 Z))$$

其中,Z_k是通道描述符,W是卷积核的宽度,$X_{i,k}$为卷积池化后的输出,W_1是编码器的权重,W_2是用于学习每通道重要性的解码器权重,outputs为通道注意力权重。

6.如权利要求1所述的预测方法,其特征在于,在训练过程中,采用基于交叉熵改进的损失函数,

$$L(\theta) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i) + \lambda\left(||\theta||^2 + \sum_{h=1}^{3}\sum_{k=1}^{16}\left\|F_{k,:,:}^{(h,1)}\right\|_2\right)$$

其中,y_i是RNA数据真实的标签,$\hat{y}_i$是经过网络预测得到的标签,$F_{k,:,:}^{(h,1)}$是后3个分支的第1个卷积层的第k个通道的卷积核,λ是正则化参数。

7.如权利要求1所述的预测方法,其特征在于,在S5中预测RNA结合位点的结构时,选取

第一个卷积层的输出中大于最大值的80％作为结合位点的潜在位点，统计该权重对应到源RNA序列的排布情况，得到不同位置上不同碱基的概率大小，构成位置权重矩阵，即结合位点的预测，该权重为第一个卷积层的输出中大于最大值的80％的输出的值。

8.根据权利要求1所述的预测方法，其特征在于，S5中在预测RNA结合位点是否存在时，卷积神经网络输出的结果是一个N*2的矩阵，每个RNA数据对应一个2维向量，向量中的2个数之和为1，表示RNA中是否存在结合蛋白的结合位点的概率大小。

9.一种基于多尺度注意力卷积神经网络的RNA结合蛋白预测装置，其特征在于，包括：

预处理模块，用于获取RNA数据并进行预处理；

编码模块，用于对预处理后的RNA数据进行编码，构建网络训练样本；

网络构建模块，用于构建多尺度注意力卷积神经网络，其中，多尺度注意力卷积神经网络包括多个分支，每个分支设置不同大小的卷积核，分别用以学习在RNA数据中的不同尺度的特征，并引入通道注意力机制学习不同通道在分类时的重要性，在进行RNA结合位点识别时，不同的通道的卷积核对应不同的结合位点结构；

网络训练模块，用于将网络训练样本输入构建的多尺度注意力卷积神经网络中，并采用Adam优化方法训练多尺度注意力卷积神经网络中的参数，得到训练好的多尺度注意力卷积神经网络；

预测模块，用于将待预测的RNA数据进行预处理和编码后输入训练好的多尺度注意力卷积神经网络中，得到预测结果，其中，预测结果包括RNA数据是否有结合蛋白对应的结合位点；

其中，网络构建模块中构建的多尺度注意力卷积神经网络包括四个分支，第一个分支包括卷积、池化、相乘、卷积、池化和相乘，第二个分支、第三个分支以及第四个分支均包括卷积、池化、卷积和池化，第一个分支的第一个相乘为第一次卷积池化后的输出结果与各通道注意力权重相乘，第二相乘为第二卷积池化后的输出结果与各通道注意力权重相乘；每个分支提取出的不同尺度的特征相加后，通过一个全连接层，得到最终预测结果。

一种基于多尺度注意力卷积神经网络的RNA结合蛋白预测方法及装置

技术领域

[0001] 本发明涉及生物信息技术领域，具体涉及一种基于多尺度注意力卷积神经网络的RNA结合蛋白预测方法及装置。

背景技术

[0002] 生物信息技术是利用数学模型、统计学方法和计算机处理生物学数据的技术，生物信息学是一门随着人类基因组计划的启动而兴起的一门新的交叉学科。在生物信息学中，对于DNA/RNA和蛋白质的研究尤其重要，DNA/RNA是生物体中遗传信息的载体和传递者，参与了遗传信息的转录翻译等重要的生化过程，蛋白质则是生命的物质基础，这种有机大分子，是构成细胞的基本有机物，是生命活动的主要承担者。研究DNA/RNA和蛋白质对于理解生命体内部的反应过程，治疗疾病等有着非常重大的意义和价值，DNA/RNA和蛋白质不仅仅各自发挥着作用，它们的相互作用调控生物体内部的反应过程，而与RNA结合的蛋白质即RNA结合蛋白。

[0003] RNA结合蛋白(RBP)在活细胞的多个生物学过程中起着重要作用，例如基因调控和mRNA定位等。基因调节包括在活生物体中大量的共转录和转录后基因表达，包括聚腺苷酸化，RNA剪接，修饰，加帽，定位，翻译和更新。研究人员发现，许多RBP的突变可能引起某些重要的疾病，例如神经退行性疾病，癌症和心血管疾病，这是由某些RBP的功能障碍引起的。因此，在这方面的深入研究可以帮助人们进一步了解许多生物学机制和相关疾病的治疗。

[0004] 高通量技术的发展极大地促进了RNA-蛋白质相互作用的基因组研究。这些高通量技术，例如交联免疫沉淀与高通量测序(CLIP-seq)，可提供大量实验验证的RBP结合位点数据。但是它仍然有一些缺点，可能需要通过一些计算方法来弥补。首先，高通量技术既费时又昂贵。其次，由于实验噪声和当前的局限性，收集到的数据中存在许多假阳性和假阴性样本。

[0005] 预测RNA中是否存在结合蛋白的结合位点这个问题是一个二分类的问题，是在给定RNA序列数据的情况下，通过学习RNA结合位点的结构，从而预测RNA数据中是否存在对应结合位点。目前，相关的方法主要用于分析DNA/RNA数据的特点以及寻找一些基因缺陷导致的疾病的病理等等。

[0006] 为了解决这些问题，国内外的科学家已提出了许多机器学习算法和计算工具来预测RBP结合位点并生成对应结合位点的结构。例如，BioBayesNet是第一个考虑结构特征，以解决转录因子结合位点的目标识别问题的工具。RNAContext是一种基序发现方法，可确定RBP对RNA序列和结构的相对结合偏好。GraphProt通过图形编码从序列和结构信息中提取大量特征，并使用支持向量机(SVM)来预测RNA结合位点是否存在。RNAcommender分析蛋白质结构域和预测的RNA二级结构，使用更高维的信息辅助从而得到更精确的预测。

[0007] 本申请发明人在实施本发明的过程中，发现现有技术的方法，至少存在如下技术问题：

[0008] 但是,这些传统的机器学习方法并不能充分提取RNA数据的潜在复杂的特征,所以导致其预测精度普遍不高。

[0009] 由此可知,现有技术中的方法存在预测精度不高的技术问题。

发明内容

[0010] 本发明提出一种基于多尺度注意力卷积神经网络的RNA结合蛋白预测方法及装置,用于解决或者至少部分解决现有技术中的方法存在的预测精度不高的技术问题。

[0011] 为了解决上述技术问题,本发明第一方面提供了一种基于多尺度注意力卷积神经网络的RNA结合蛋白预测方法,包括:

[0012] S1:获取RNA数据并进行预处理;

[0013] S2:对预处理后的RNA数据进行编码,构建网络训练样本;

[0014] S3:构建多尺度注意力卷积神经网络,其中,多尺度注意力卷积神经网络包括多个分支,每个分支设置不同大小的卷积核,分别用以学习在RNA数据中的不同尺度的特征,并引入通道注意力机制学习不同通道在分类时的重要性,在进行RNA结合位点识别时,不同的通道的卷积核对应不同的结合位点结构;

[0015] S4:将网络训练样本输入构建的多尺度注意力卷积神经网络中,并采用Adam优化方法训练多尺度注意力卷积神经网络中的参数,得到训练好的多尺度注意力卷积神经网络;

[0016] S5:将待预测的RNA数据进行预处理和编码后输入训练好的多尺度注意力卷积神经网络中,得到预测结果,其中,预测结果包括RNA数据是否有结合蛋白对应的结合位点。

[0017] 在一种实施方式中,S1具体包括:将获取的不同长度的RNA数据填补至相同的长度。

[0018] 在一种实施方式中,S2具体包括:

[0019] 将预处理后的RNA数据采用One-hot矩阵表示,构成网络训练样本。

[0020] 在一种实施方式中,S3中构建的多尺度注意力卷积神经网络包括四个分支,第一个分支包括卷积、池化、相乘、卷积、池化和相乘,第二个分支、第三个分支以及第四个分支均包括卷积、池化、卷积和池化,第一个分支的第一个相乘为第一次卷积池化后的输出结果与各通道注意力权重相乘,第二相乘为第二卷积池化后的输出结果与各通道注意力权重相乘;每个分支提取出的不同尺度的特征相加后,通过一个全连接层,得到最终预测结果。

[0021] 在一种实施方式中,构建的网络训练样本包括正训练样本和负训练样本,蛋白质对应的正训练样本为包含有该蛋白质结合位点的RNA数据,负训练样本为无该蛋白质结合位点的RNA数据,训练过程中,正训练样本标签赋值为1,负训练样本标签赋值为0。

[0022] 在一种实施方式中,通道注意力权重的计算方式为:

[0023] $$Z_k = \frac{1}{W}\sum_{i=1}^{W} X_{i,k}$$

[0024] $outputs = softmax(W_2 sigmoid(W_1 Z))$

[0025] 其中,Z_k是通道描述符,W是卷积核的宽度,$X_{i,k}$为卷积池化后的输出,W_1是编码器的权重,W_2是用于学习每通道重要性的解码器权重,outputs为通道注意力权重。

[0026] 在一种实施方式中，在训练过程中，采用基于交叉熵改进的损失函数，

[0027] $$L(\theta) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i) + \lambda(||\theta||^2 + \sum_{h=1}^{3}\sum_{k=1}^{16}\left\|F_{k,:,:}^{(h,1)}\right\|_2)$$

[0028] 其中，y_i是RNA数据真实的标签，$\hat{y}_i$是经过网络预测得到的标签，$F_{k,:,:}^{(h,1)}$是后3个分支的第1个卷积层的第k个通道的卷积核，λ是正则化参数。

[0029] 在一种实施方式中，在S5中预测RNA结合位点的结构时，选取第一个卷积层的输出中大于最大值的80％作为结合位点的潜在位点，统计该权重对应到源RNA序列的排布情况，得到不同位置上不同碱基的概率大小，构成位置权重矩阵，即结合位点的预测。

[0030] 在一种实施方式中，S5中在预测RNA结合位点是否存在时，卷积神经网络输出的结果是一个N*2的矩阵，每个RNA数据对应一个2维向量，向量中的2个数之和为1，表示RNA中是否存在结合蛋白的结合位点的概率大小。

[0031] 基于同样的发明构思，本发明第二方面提供了一种基于多尺度注意力卷积神经网络的RNA结合蛋白预测装置，包括：

[0032] 预处理模块，用于获取RNA数据并进行预处理；

[0033] 编码模块，用于对预处理后的RNA数据进行编码，构建网络训练样本；

[0034] 网络构建模块，用于构建多尺度注意力卷积神经网络，其中，多尺度注意力卷积神经网络包括多个分支，每个分支设置不同大小的卷积核，分别用以学习在RNA数据中的不同尺度的特征，并引入通道注意力机制学习不同通道在分类时的重要性，在进行RNA结合位点识别时，不同的通道的卷积核对应不同的结合位点结构；

[0035] 网络训练模块，用于将网络训练样本输入构建的多尺度注意力卷积神经网络中，并采用Adam优化方法训练多尺度注意力卷积神经网络中的参数，得到训练好的多尺度注意力卷积神经网络；

[0036] 预测模块，用于将待预测的RNA数据进行预处理和编码后输入训练好的多尺度注意力卷积神经网络中，得到预测结果，其中，预测结果包括RNA数据是否有结合蛋白对应的结合位点。

[0037] 本申请实施例中的上述一个或多个技术方案，至少具有如下一种或多种技术效果：

[0038] 本发明提供的一种基于多尺度注意力卷积神经网络的RNA结合蛋白预测方法，采用的多尺度注意力卷积神经网络包括多个分支，每个分支设置不同大小的卷积核，分别用以学习在RNA数据中的不同尺度的特征，能更大程度的提取RNA数据中的有用特征，提升了模型的鲁棒性，显著提升了在数据量较少的集合蛋白对应的RNA数据上的预测精度。

[0039] 进一步地，引入了通道注意力机制，通过输出第一个卷积层中不同通道的重要性权重，使每个通道对应的参数收敛到对RNA数据分类最重要的形式，这样增大了模型预测RNA结合位点结构的精度。

[0040] 进一步地，本发明在基于多尺度注意力卷积神经网络中提出了一种基于交叉熵改进的损失函数，加快了模型收敛的速度。提升了模型的泛化能力，从而提升目标检测的效果。

附图说明

[0041] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0042] 图1是本发明提供的一种基于多尺度注意力卷积神经网络的RNA结合蛋白预测方法的流程图；

[0043] 图2是本发明实施例中构建的多尺度注意力卷积神经网络结构示意图。

具体实施方式

[0044] 本发明提出了一种基于多尺度注意力卷积神经网络的RNA结合蛋白预测方法，多尺度注意力卷积神经网络包括多个分支，每个分支设置不同大小的卷积核，分别用以学习在RNA数据中的不同尺度的特征，并引入通道注意力机制学习不同通道在分类时的重要性，使每个通道对应的参数收敛到对RNA数据分类最重要的形式，这样增大了模型预测RNA结合位点结构的精度。

[0045] 本发明的技术方案是：

[0046] 本发明公开了一种基于多尺度注意力卷积神经网络的预测RNA结合蛋白方法，包括训练阶段和预测阶段。训练阶段包括RNA数据的预处理，RNA数据的编码，构建神经网络和网络参数训练。本发明将RNA的数学抽象的统计模式转化成矩阵的形式，输入到预先设计好的多尺度注意力卷积神经网络，通过使本发明设计的特别交叉熵损失函数最小，使用Adam优化方法训练神经网络中的参数。在预测阶段，以四个碱基为基本单元的RNA序列数据被输入到网络中，神经网络最后一层输出RNA数据中是否有结合蛋白对应的结合位点的概率大小，从而获得对RNA序列类别的预测结果。同时，通过分析网络中第一层卷积核的参数，可以统计得到结合蛋白在RNA序列数据上结合位点的结构和概率分布预测。

[0047] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0048] 实施例一

[0049] 本实施例提供了一种基于多尺度注意力卷积神经网络的RNA结合蛋白预测方法，该方法包括：

[0050] S1：获取RNA数据并进行预处理；

[0051] S2：对预处理后的RNA数据进行编码，构建网络训练样本；

[0052] S3：构建多尺度注意力卷积神经网络，其中，多尺度注意力卷积神经网络包括多个分支，每个分支设置不同大小的卷积核，分别用以学习在RNA数据中的不同尺度的特征，并引入通道注意力机制学习不同通道在分类时的重要性，在进行RNA结合位点识别时，不同的通道的卷积核对应不同的结合位点结构；

[0053] S4：将网络训练样本输入构建的多尺度注意力卷积神经网络中，并采用Adam优化方法训练多尺度注意力卷积神经网络中的参数，得到训练好的多尺度注意力卷积神经网

络；

[0054] S5：将待预测的RNA数据进行预处理和编码后输入训练好的多尺度注意力卷积神经网络中，得到预测结果，其中，预测结果包括RNA数据是否有结合蛋白对应的结合位点。

[0055] 现有的一些基于深度的方法也运用到了RNA结合蛋白预测当中来。由于不同的结合蛋白对应的数据数量不同，深度学习方法在不同的数据上得到的结果差异性很大。在数据量较大、数据多样性高的数据上能取得很好的效果；而在数据量较小、数据形式单一的数据上会产生过拟合现象，导致效果普遍偏低。同时，因为生物实验得到的数据会有很多噪音，所以通过深度学习预测的RNA结合位点的结构的准确率也会受到影响。故RNA数据的高噪音和部分结合蛋白对应的RNA数据的匮乏，给在预测RNA结合蛋白中使用深度学习这样一种提取特征的强大工具也带来了困难。

[0056] 因此，本发明提供了一种基于多尺度注意力卷积神经网络的RNA结合蛋白预测方法，S1～S4为训练阶段，S5为预测阶段。图1为具体实现流程图，One-hot矩阵为编码后得到的结果，预测结果处理包括结合位点是否存在之外，还包括结合位点的结构。

[0057] 具体来说，每个分支设置不同大小的卷积核，可以增强模型表达RNA数据的能力，提高目标分类精度。在用于分类RNA数据和学习结合位点结构的神经网络中引入了通道注意力机制，学习不同通道在分类时的重要性。在识别结合位点时，不同的通道的卷积核对应不同的结合位点结构，注意力可帮助模型在优化过程中，所有的卷积核都学习到最可能的结合位点结构。

[0058] 具体实施时，采用Python平台基于TensorFlow实现，使用了TOMTOM和AME软件对预测的RNA结合蛋白的结构进行可视化处理。TOMTOM通过从RBP数据库中搜索给定的查询结合位点，将数据库中已有的RNA结合位点和想要查询的结合位点之间进行比对，推测碱基在结合位点的概率分布，从而推测对应的结合位点的类别。AME工具，可以通过针对输入序列和相应的混洗序列扫描预测的结合位点来估计富集得分，一般情况下得分越高的结合位点就越有可能是真实的结合位点。这两个工具已完全集成到MEME工具中。RNA数据集合X中N个RNA数据的长度彼此不同，需要进行预处理之后才能成为计算机可用的数据。

[0059] 在一种实施方式中，S1具体包括：将获取的不同长度的RNA数据填补至相同的长度。

[0060] 具体实施过程中，可以统计数据集合X中n个RNA序列的长度，最长的长度设为L_{max}，已知结合位点对应的碱基长度为m，将m-1个'N'填补至RNA序列前，余下的'N'填补至RNA序列后，直至整体的RNA序列的长度为L_{max}，数据集合X包括获取的RNA数据，RNA序列即RNA数据，'N'为填补的占位符。RNA在生物意义上是由4种不同的碱基构成的，RNA的每个位置是A，C，G，U这4种碱基中的某一种。

[0061] 在一种实施方式中，S2具体包括：

[0062] 将预处理后的RNA数据采用One-hot矩阵表示，构成网络训练样本。

[0063] 具体实施过程中，经过填补之后的每个RNA数据的长度为L_{max}，此时每个RNA数据的每个位置上由5个基本元素'A'，'C'，'G'，'U'，'N'构成，其中前4个基本元素对应RNA中4种不同的碱基类别，'N'是填补位置的占位符。对于给定的未填补的RNA序列数据$s=\{s_1, s_2, \ldots, s_n\}$，按照以下的方式将填补后的RNA数据转换为One-hot矩阵：

[0064] $$M_{i,:}=\begin{cases}\left[\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right] \ if\ s_{i-m+1}\ is\ 'N'\\ [1,0,0,0]\ if\ s_{i-m+1}\ is\ 'A'\\ [0,1,0,0]\ if\ s_{i-m+1}\ is\ 'C'\\ [0,0,1,0]\ if\ s_{i-m+1}\ is\ 'G'\\ [0,0,0,1]\ if\ s_{i-m+1}\ is\ 'U'\end{cases}$$

[0065] 按照上述的变换法则,得到的One-hot矩阵的大小为$L_{max}\times 4$,这样的固定大小的数据即可输入到神经网络中进行训练或者预测。

[0066] One-hot矩阵处理方式即每个碱基变为一个4维向量,真实碱基对应的维度的值为1,否则为0。举例来说,将不同长度One-hot矩阵填补成同样长度,使用4个碱基平均分布的方式进行填补,即One-hot矩阵中每个填补的位置对应的4维向量为(0.25,0.25,0.25,0.25)。由于结合蛋白对应的RNA上的结合位点的长度一般为7,RNA数据前填补6位,RNA数据后填补至所有RNA数据的最大长度。

[0067] 在一种实施方式中,S3中构建的多尺度注意力卷积神经网络包括四个分支,第一个分支包括卷积、池化、相乘、卷积、池化和相乘,第二个分支、第三个分支以及第四个分支均包括卷积、池化、卷积和池化,第一个分支的第一个相乘为第一次卷积池化后的输出结果与各通道注意力权重相乘,第二相乘为第二卷积池化后的输出结果与各通道注意力权重相乘;每个分支提取出的不同尺度的特征相加后,通过一个全连接层,得到最终预测结果。

[0068] 具体来说,请参见图2,为一种实施方式中的网络结构图,该实施方式中,网络包括4个分支,每个分支具有不同大小的卷积核,用以提取不同尺度的特征。

[0069] 其中,第一个分支中引入了通道注意力机制,其中,保留第一次卷积、池化后的结果,同时,将第一次卷积池化后的结果通过全局池化、两个全连接层和Softmax激活函数,得到通道注意力权重,然后与第一次卷积池化后的结果相乘,进行后续操作(即用于优化之后的结合位点预测),第二卷积池化、相乘的实现过程与前述过程类似,在此不再详述。

[0070] 在一种实施方式中,构建的网络训练样本包括正训练样本和负训练样本,蛋白质对应的正训练样本为包含有该蛋白质结合位点的RNA数据,负训练样本为无该蛋白质结合位点的RNA数据,训练过程中,正训练样本标签赋值为1,负训练样本标签赋值为0。

[0071] 具体来说,正负训练样本的数量由已有的数据库中对应数据的多少而定,故不同的蛋白质对应的RNA序列数据存在差异。

[0072] 在一种实施方式中,通道注意力权重的计算方式为:

[0073] $$Z_k=\frac{1}{W}\sum_{i=1}^{W}X_{i,k}$$

[0074] $outputs=softmax(W_2 sigmoid(W_1 Z))$

[0075] 其中,Z_k是通道描述符,W是卷积核的宽度,$X_{i,k}$为卷积池化后的输出,W_1是编码器的权重,W_2是用于学习每通道重要性的解码器权重,outputs为通道注意力权重。

[0076] 具体来说,训练过程中优选建议网络批训练数目设为512,网络学习率设为1.0e-3,正则化参数设为1.0e-3,神经元的丢弃率设为0.25。其中的多尺度特点在于在整体的网络结构中设置了不同的分支,各个分支的结构类似,其中的卷积核的大小彼此有差异,不同的分支用于提取不同尺度的RNA特征。通道注意力嵌入到网络中体现在第一个分支中利用

自编码机学习到各个通道对应的重要性权重，从而保证之后每个通道对应卷积核都能收敛到对分类重要的参数形式。

[0077] 这里outputs即为各个通道对应的重要性权重，这里的outputs是attention层的outputs，而attention层是用于得到各个通道对应的重要性权重，故这里用outputs指代。

[0078] 在图2所示的网络模型中，有16个通道，每个通道对应一个位置权重矩阵，即16种位置权重矩阵，该矩阵用于提取RNA序列数据中的结合位点。重要性权重是模型学习到的每个通道重要性大小，用于之后的模型预测优化。通道对应的重要性权重即衡量位置权重矩阵的重要性权重。

[0079] 在一种实施方式中，在训练过程中，采用基于交叉熵改进的损失函数，

[0080] $$L(\theta) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i) + \lambda(||\theta||^2 + \sum_{h=1}^{3}\sum_{k=1}^{16}\left\|F_{k,:,:}^{(h,1)}\right\|_2)$$

[0081] 其中，y_i是RNA数据真实的标签，$\hat{y}_i$是经过网络预测得到的标签，$F_{k,:,:}^{(h,1)}$是后3个分支的第1个卷积层的第k个通道的卷积核，λ是正则化参数。

[0082] 具体来说，对交叉熵损失函数进行了优化，在原有的损失函数中加上了第一层卷积的L2范数，这可以帮助模型学习到更好的特征，预测更加准确。

[0083] 在一种实施方式中，在S5中预测RNA结合位点的结构时，选取第一个卷积层的输出中大于最大值的80％作为结合位点的潜在位点，统计该权重对应到源RNA序列的排布情况，得到不同位置上不同碱基的概率大小，构成位置权重矩阵，即结合位点的预测。

[0084] 具体来说，已知第一层卷积池化的输出(即第一个分支第一次卷积池化后的输出)，该输出经过激活函数得到输出Z，每个通道都会对应有一个输出Z，统计不同位置上得到Z的值的大小。其中，一个卷积操作会有不同的通道，每个通道对应一种位置权重矩阵，一种位置权重矩阵为一个通道，RNA序列数据经过卷积操作之后，会得到不同通道的输出，该输出越大，表明原RNA序列对应位置的RNA片段和位置权重矩阵越相似。每个分支中都会有卷积操作，故一个分支包含多个通道。

[0085] 因此，对每个通道，输出值大于最大值的80％的位置，预测是存在结合位点的。若该位置为i，则预测RNA序列s中$S_{i-\frac{k}{2}}, S_{i-\frac{k}{2}+1}, \ldots, S_{i+\frac{k}{2}}$为可能的结合位点，最终将这个统计数据输入到TOMTOM工具中，可视化预测的结合位点的概率分布情况。

[0086] 举例来说，若卷积核为4*K，卷积操作后，原序列中长度为K的片段变为一个值，位置i的可能的结合位点就是卷积操作前RNA序列中位置i的长度为K的RNA片段。统计原RNA序列中的长度为K的片段和其卷积计算后的权重大小。然后截取最大的20％权重对应的长度为K的RNA片段，计算其中每个碱基出现的概率大小，从而得到概率分布预测。

[0087] 在一种实施方式中，S5中在预测RNA结合位点是否存在时，卷积神经网络输出的结果是一个N*2的矩阵，每个RNA数据对应一个2维向量，向量中的2个数之和为1，表示RNA中是否存在结合蛋白的结合位点的概率大小。

[0088] 本发明提供的预测方法，在具体实施时，可采用软件方式实现流程的自动运行。运行流程的装置也应当在本发明的保护范围内。

[0089] 以下通过对比试验来验证本发明的有益效果。

[0090] 本试验采用的数据从HITS-CLIP,PAR-CLIP,iCLIP这3个数据库中提取而得,该数据一共包括24种结合蛋白,分别是Ago1-4,IGF2BP1-3,ZC3H7B,TIAL1,TIA1,TDP-43,TAF15,SFRS1,QKI,PUM2,PTB,MOV10,HNRNPC,FUS,EWSR1,CAPRIN1,C22ORF28,C17ORF85,ALKBH5。每个RNA数据的长度在200-500之间,不同的蛋白质对应的RNA数据的个数差异很大。分别采用Pse-SVM(方法1),GraphProt(方法2),Deepnet-rbp(方法3),iDeepE(方法4)和本发明方法进行预测比较,本发明方法以具体实施方式的方法为例。

[0091] RNA结合蛋白预测评价指标:AUC(ROC曲线下面积)值。

[0092] AUC值由ROC(接收器操作特性曲线)曲线下面积计算得到。根据一般预测过程,预测与阈值有关。在一定的阈值下,有些含有结合位点的RNA被正确预测出来,即为真正类(TP),有些会被漏检,也有无结合位点的RNA被预测为正类,即为假正类(FP)。因此,阈值的设置非常重要,通常需要在达到较高的真正类率的同时,保持较低的假负类率。真正类率TPR和假负类率(TNR)的定义是:

[0093] $TPR=N_{TP}/N_T$

[0094] $FPR=N_{FP}/N$

[0095] 其中N_{TP}表示在给定的阈值下检测出来的真实RNA数量,N_T表示总体RNA数据中的正样本的数量,N_{FP}表示被误分为正类的负类RNA数据,N表示的是总体RNA数据的数量。将真正类率作为纵坐标,假负类率作为横坐标即可绘制得到ROC曲线,通过积分得到曲线下面积AUC值。

[0096] 表1对比试验结果

[0097]

RBP	本发明方法	方法 1	方法 2	方法 3	方法 4
ALKBH5	0.812	0.648	0.68	0.714	0.758

[0098]

C17ORF85	0.903	0.734	0.8	0.82	0.83
C22ORF28	0.881	0.764	0.751	0.792	0.837
CAPRIN1	0.918	0.728	0.855	0.834	0.893
Ago2	0.895	0.746	0.765	0.809	0.884
ELAVL1H	0.982	0.816	0.955	0.966	0.979
SFRS1	0.955	0.746	0.898	0.931	0.946
HNRNPC	0.98	0.824	0.952	0.962	0.976
TDP43	0.951	0.84	0.874	0.876	0.945
TIA1	0.954	0.784	0.861	0.891	0.937
TIAL1	0.947	0.724	0.833	0.87	0.934
Ago1-4	0.928	0.728	0.895	0.881	0.915
ELAVL1B	0.976	0.837	0.935	0.961	0.971
ELAVL1A	0.982	0.83	0.959	0.966	0.964
EWSR1	0.976	0.753	0.935	0.966	0.969
FUS	0.987	0.762	0.968	0.98	0.985
ELAVL1C	0.992	0.853	0.991	0.994	0.988
IGF2BP1-3	0.967	0.753	0.889	0.879	0.947
MOV10	0.941	0.783	0.863	0.854	0.916
PUM2	0.978	0.84	0.954	0.971	0.967
QKI	0.981	0.809	0.957	0.983	0.97
TAF15	0.988	0.769	0.97	0.983	0.976
PTB	0.951	0.867	0.937	0.983	0.944
ZC3H7B	0.912	0.743	0.82	0.796	0.907
Mean	0.947375	0.778	0.887	0.902	0.931

[0099] 从表1可见，本发明方法在试验的24组数据上都能获得更高的AUC值，表明本发明的方法具有更强的RNA结合蛋白的预测能力。与传统的机器学习的方法(方法1和2)相比，本发明方法的AUC值有大幅度的提高，表明本发明方法比传统机器学习方法的蛋白质预测能力强很多；而与现有的深度学习方法(如方法3和4)相比，本发明方法的AUC值也更高。对于数据量比较少的ALKBH5、C17ORF85等蛋白质对应的数据上，本发明方法的效果比所有的对比方法都有显著提升。同时通过预测得到的结合位点结构的概率分布可以发现，本发明在

预测RBP结合位点的结构和概率分布上，比现有的RNA结合蛋白预测方法的效果要好。

[0100] 由此可得出结论，与已有RNA结合蛋白预测方法相比，本发明方法拥有更高的预测精度。本发明解决了目标训练样本不足导致的在深度网络上预测准确率低下的问题，通过多尺度注意力卷积神经网络提取RNA数据多尺度的特征，有效提升了模型的鲁棒性和泛化能力。本发明在神经网络中引入了通道注意力的机制，选择最重要的候选卷积核用于提取可能的结合位点，提升了预测结合位点结果的准确率。同时，本发明改进了神经网络的损失函数，使神经网络能在更一般的数据上取得更好的效果。

[0101] 实施例二

[0102] 基于同样的发明构思，本发明第二方面提供了一种基于多尺度注意力卷积神经网络的RNA结合蛋白预测装置，该装置包括：

[0103] 预处理模块，用于获取RNA数据并进行预处理；

[0104] 编码模块，用于对预处理后的RNA数据进行编码，构建网络训练样本；

[0105] 网络构建模块，用于构建多尺度注意力卷积神经网络，其中，多尺度注意力卷积神经网络包括多个分支，每个分支设置不同大小的卷积核，分别用以学习在RNA数据中的不同尺度的特征，并引入通道注意力机制学习不同通道在分类时的重要性，在进行RNA结合位点识别时，不同的通道的卷积核对应不同的结合位点结构；

[0106] 网络训练模块，用于将网络训练样本输入构建的多尺度注意力卷积神经网络中，并采用Adam优化方法训练多尺度注意力卷积神经网络中的参数，得到训练好的多尺度注意力卷积神经网络；

[0107] 预测模块，用于将待预测的RNA数据进行预处理和编码后输入训练好的多尺度注意力卷积神经网络中，得到预测结果，其中，预测结果包括RNA数据是否有结合蛋白对应的结合位点。

[0108] 由于本发明实施例二所介绍的装置，为实施本发明实施例一中基于多尺度注意力卷积神经网络的RNA结合蛋白预测方法所采用的装置，故而基于本发明实施例一所介绍的方法，本领域所属人员能够了解该装置的具体结构及变形，故而在此不再赘述。凡是本发明实施例一的方法所采用的装置都属于本发明所欲保护的范围。

[0109] 本领域内的技术人员应明白，本发明的实施例可提供为方法、系统、或计算机程序产品。因此，本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0110] 本发明是参照根据本发明实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0111] 尽管已描述了本发明的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例做出另外的变更和修改。所以，所附权利要求意欲解释为包括优

选实施例以及落入本发明范围的所有变更和修改。

[0112] 显然,本领域的技术人员可以对本发明实施例进行各种改动和变型而不脱离本发明实施例的精神和范围。这样,倘若本发明实施例的这些修改和变型属于本发明权利要求及其等同技术的范围之内,则本发明也意图包含这些改动和变型在内。

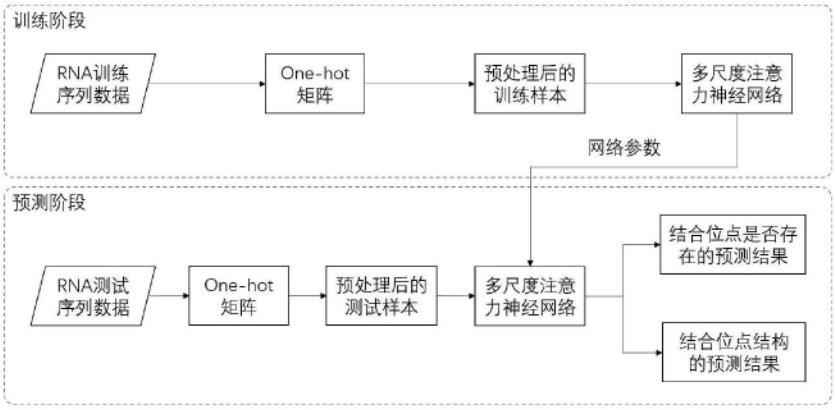

训练阶段
RNA训练序列数据
One-hot矩阵
预处理后的训练样本
多尺度注意力神经网络
网络参数
预测阶段
RNA测试序列数据
One-hot矩阵
预处理后的测试样本
多尺度注意力神经网络
结合位点是否存在的预测结果
结合位点结构的预测结果

图1

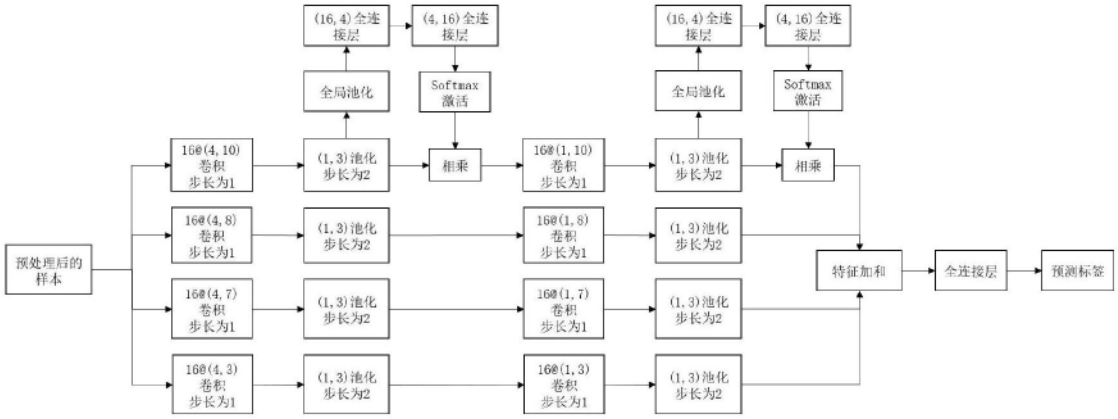

预处理后的样本
16@(4,10)卷积步长为1
16@(4,8)卷积步长为1
16@(4,7)卷积步长为1
16@(4,3)卷积步长为1
(1,3)池化步长为2
(1,3)池化步长为2
(1,3)池化步长为2
(1,3)池化步长为2
全局池化
Softmax激活
相乘
(16,4)全连接层
(4,16)全连接层
16@(1,10)卷积步长为1
16@(1,8)卷积步长为1
16@(1,7)卷积步长为1
16@(1,3)卷积步长为1
(1,3)池化步长为2
(1,3)池化步长为2
(1,3)池化步长为2
(1,3)池化步长为2
全局池化
Softmax激活
相乘
(16,4)全连接层
(4,16)全连接层
特征加和
全连接层
预测标签

图2