Hans Journal of Computational Biology 计算生物学, 2022, 12(2), 9-15
Published Online June 2022 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2022.122002

Hans汉斯

基于机器学习的RNA甲基化修饰位点预测的研究进展

纪璎珊

辽宁科技大学计算机与软件工程学院，辽宁 鞍山

收稿日期：2022年4月30日；录用日期：2022年5月30日；发布日期：2022年6月9日

摘　要

RNA修饰，特别是RNA甲基化，在人类多种生物活动中起着非常重要的调控作用，最常见的修饰包括N6-腺苷酸甲基化(m6A)、N1-腺苷酸甲基化(m1A)、胞嘧啶羟基化(m5C)等。RNA甲基化修饰位点的准确识别对预测多种人类遗传学疾病以及药物研发发挥着关键作用。随着数据集的大量积累，序列数据的分析需求不断增多，一些基于机器学习的预测方法被开发出来，用于甲基化位点的识别。本工作分别从RNA修饰、数据集来源、预测结果的评估标准以及用于预测的算法模型优缺点等方面进行综述，最后指出了RNA甲基化修饰位点预测未来的研究方向。

关键词

RNA甲基化，位点预测，特征分析，机器学习

Research Progress of RNA Methylation Modification Site Prediction Based on Machine Learning

Yingshan Ji

School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan Liaoning

Received: Apr. 30th, 2022; accepted: May 30th, 2022; published: Jun. 9th, 2022

Abstract

RNA modification, especially RNA methylation, plays a very important regulatory role in a variety

文章引用: 纪璎珊. 基于机器学习的 RNA 甲基化修饰位点预测的研究进展[J]. 计算生物学, 2022, 12(2): 9-15.
DOI: 10.12677/hjcb.2022.122002

of human biological activities. The most common modifications include N6-adenylate methylation (m6A), N1-adenylate methylation (m1A), cytosine hydroxylation (m5C), etc. Accurate identification of RNA methylation modification sites is crucial for predicting a variety of human genetic diseases and drug development. With the accumulation of a large number of data sets, the requirements of analyzing sequence data are increasing, and some prediction methods based on machine learning have been developed for the identification of methylation sites. This work reviews RNA modification, data set sources, evaluation criteria for prediction results, and advantages and disadvantages of algorithm models used for prediction, and finally presents the research direction of RNA methylation modification site prediction in the future.

Keywords

RNA Methylation, Site Prediction, Feature Analysis, Machine Learning

1. 引言

RNA 修饰是指真核生物和原核生物中 RNA 的转录后修饰。目前，超过 100 种不同类型的 RNA 修饰已在所有生物体中进行了表征。RNA 修饰发生在多种 RNA 分子中，包括 mRNA、tRNA、rRNA、lncRNA 和 snoRNA，在 RNA 剪接、蛋白质定位和翻译、干细胞多能性和人类疾病中发挥着重要作用。mRNA 中最常见的内部修饰包括 N6-腺苷酸甲基化(m6A)、N1-腺苷酸甲基化(m1A)、胞嘧啶羟基化(m5C)等。其中最主要的是 RNA 甲基化，通常被称为表观转录组[1]。

m6A 是 6 位氮的甲基化腺苷，发生在 mRNA 加工、核输出、翻译调控及 RNA 降解的不同阶段，包括 ncRNA 加工和 CircRNA 翻译。据估计，m6A 甲基化大约存在于四分之一的 mRNA 上。多项研究证明 m6A 修饰是动态可逆的，能够起到促进环状 RNA 翻译、通过促进 mRNA 降解来调控癌症干细胞的分化，以及调控 T 细胞分化及免疫稳态等作用[2]。m1A 普遍存在于非编码 RNA 和 mRNA 中，是 RNA 分子腺嘌呤第 1 位氮原子上的甲基化修饰，研究表明，m1A 与呼吸链功能障碍和神经发育退化有关[3]。影响 RNA 胞嘧啶碱基的修饰主要包括 m5C、5hmC 等。m5C 被定义为甲基在胞嘧啶的第五个碳原子上的加入，存在于多种 RNA 中。最近的研究表明，m5C 甲基化可促进 mRNA 的转运，提高核质穿梭的效率，并对 mRNA 稳定、胚胎发生和肿瘤发生产生积极影响[4]。5-羟甲基胞嘧啶(5hmC)则是 TET 介导的 m5C 氧化产生另一种形式的 RNA 修饰。

RNA 修饰位点的识别主要基于生化实验检测或计算预测，但随着数据集的大量积累，便突出了生化检测高成本且耗时的缺陷，由此，机器学习算法逐步在 RNA 修饰预测的领域崭露头角。本文介绍了几种 RNA 甲基化研究的常用数据集，并就常见的 RNA 甲基化位点介绍几种基于机器学习的预测方法，根据评估标准对比模型之间的性能优势。

2. 基准数据集

训练高效计算模型的一个重要步骤是构建高质量的数据集。在 RNA 修饰的研究中，基准数据集大多来源于开源数据库 Gene Expression Omnibus (GEO) [5]。GEO 是 2000 年由美国国立生物技术信息中心创建，收录了世界各国研究机构所提交的高通量基因表达数据，通过限定检测类型，如：DNA、mRNA、

甲基化等检索具体数据。另一种常用的数据库是 RMBase，一个整合了表观转录组测序数据的综合数据库，该数据库由屈良鹄教授实验室构建，并于 2017 年更新了 RMBase V2.0 [6]，与之前的版本相比，增加了大量的 RNA 修饰位点数据。这些数据库为基于机器学习的多种模型方法提供基准数据集，训练数据的质量对于模型的预测效果的影响远超模型的选择与构建。

3. 模型性能评估

采用四种性能指标评估模型的性能，即 S_n (灵敏度)、S_p (特异性)、ACC (准确性)、MCC (马修斯的相关系数)。在这些指标中，S_n 表示该模型在预测阳性样本方面的准确性。Sn 越高，说明对阳性样本的预测性能较高。同时，S_p 越高，说明对阴性样本的预测性能越高。ACC 代表了真阳性和真阴性样本预测的成功率。一个好的预测模型应该同时具有高 S_n 和 S_p。如果 S_n 很高，S_p 很低，则产生高假阳性，而如果 S_p 很高，S_n 很低，则产生高假阴性。此外，MCC 可以反映结果的可靠性，这对样本比例的不平衡是稳健的。这四个指标的定义如下

$$S_n = \frac{TP}{TP + FN} \tag{1}$$

$$S_P = \frac{TN}{TN + FP} \tag{2}$$

$$ACC = \frac{TP + TN}{TN + FP + TP + FN} \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}} \tag{4}$$

其中，TP、TN、FP、FN 分别为真阳性、真阴性、假阳性、假阴性等值。此外，还使用曲线下面积(AUC)作为评估模型预测性能的有效指标。

4. 预测模型

4.1. N6-Methyladenosine

N6-甲基腺苷(m6A)是一种典型且广泛的转录后 RNA 修饰，几乎影响所有细胞周期过程，早期发现后，通过高通量实验从不同物种中鉴定出数百或数千个 m6A 位点，为构建 m6A 位点识别的计算机方法研究提供了丰富的数据集资源。数据集的充足使预测 m6A 位点的方法逐步趋于成熟。现有的 m6A 预测器主要是使用传统的机器学习算法开发的。

iRNA-Methyl [7]是第一个使用机器学习方法进行 m6A 位点识别的开创性研究，该模型由 Chen 等人构建，使用 SVM 模型。第二年 Zhou 等人提出了一个名为“SRAMP”预测工具[8]。在他们提出的方法中，使用了多种类型的特征描述符，包括核苷酸序列的位置二进制编码、k-最近邻编码、核苷酸对频谱编码和二级结构模式，用于训练基于随机森林的 m6A 集成预测模型。与其他现有预测器相比，他们提出的方法取得了相对更好的性能。M6AMRFS [9]是一种基于序列的预测器，用于检测多个物种的 RNA 序列中的 m6A 位点。他们通过使用二核苷酸二进制编码和局部位置特异性二核苷酸频率对序列进行编码，提出了一种特征表示算法。他们将 F-score 算法与顺序前向搜索相结合，以优化特征空间并提高表示能力。他们采用 XGBoost 算法对可用的最佳特征执行模型训练。

近年来，除了传统的机器学习算法，深度学习已成为一种流行且强大的工具，因为它提供了多层网络和非线性映射操作，以数据驱动的方式检测潜在的复杂模式。深度学习方法在解决几个预测问题，如 RNA 剪接、蛋白质结构和蛋白质修饰等方面已经证明了优于传统机器学习算法的性能。Nazari 等人则提

出了一种基于卷积神经网络(CNN)的 m6A 预测模型，名为 iN6-Methy (5-step) [10]，用于 *H. sapiens*、*M. musculus* 和 *S. cerevisiae* 基准物种的 m6A 位点预测。在他们提出的方法中，他们使用基于自然语言处理的 word2vec 模型提取特征。在这种方法中，使用 k-mer 技术将每个序列手动分割成长度为 k 的序列段。他们将 k 的值设置为 3，并将每个序列段映射到其对应的特征表示。由于模型使用整个基因组进行训练，其计算复杂度很高，而对 m6A 位点的预测速度很慢。2020 年，Alam 等人提出了 pm6A-CNN [11]模型，使用 one-hot 编码和核苷酸化学特性(NCP)的组合作为模型的输入，卷积神经网络作为分类方法。此外，该模型使用网格搜索算法来确定模型的最佳参数。与现有方法相比，他们提出的方法实现了改进的性能。2021 年，M6A-NeuralTool [12]模型使用三个子体系结构来预测 N6-甲基腺苷位点的修饰，三个子体系结构分别使用完全连通层、支持向量机和朴素贝叶斯进行分类。目前，性能优于现有用于 m6A 位点识别的模型。M6A 位点预测工具性能总结，如表 1 所示。

Table 1. Performance of the M6A modification site prediction tool
表 1. M6A 修饰位点预测工具的性能

Method	Species	ML-Algorithm	Sn	Sp	ACC	MCC
iRNA-Methyl	*S. cerevisiae*	SVM	0.706	0.606	0.656	0.290
M6AMRFS	*S. cerevisiae*	XGBoost	0.752	0.733	0.743	0.099
	A. thaliana		0.807	0.814	0.811	0.621
	M. musculus		0.828	0.758	0.793	0.758
	H. sapiens		0.820	1.000	0.910	0.833
iN6-Methyl(5-step)	*S. cerevisiae*	CNN	0.762	0.746	0.754	0.507
	M. musculus		0.789	1.000	0.895	0.807
	H. sapiens		0.821	1.000	0.911	0.835
pm6A-CNN	*S. cerevisiae*	CNN	0.846	0.855	0.850	0.703
	A. thaliana		0.923	0.926	0.925	0.850
	M. musculus		0.904	0.972	0.938	0.880
	H. sapiens		0.886	0.986	0.936	0.878
M6A-NeuralTool	*S. cerevisiae*	CNN	0.715	0.716	0.715	0.466
	A. thaliana		0.939	0.944	0.942	0.872
	M. musculus		0.915	1.000	0.958	0.912
	H. sapiens		0.920	1.000	0.960	0.882

4.2. 5-Methylcytosine

确定 m5C 位点在 RNA 中的位置对于理解转录后修饰的机制和功能至关重要，而传统鉴定 m5C 的高通量测序方法当面临大量待测数据时，需要花费大量时间与实验成本，大大影响了检测效率。近年来，已然发展了一些用于识别 M5C 位点的机器学习方法。M5C-PseDNC [13]是第一个用于 m5C 位点预测的模型，采用 PseDNC 来构造样本，经过特征提取再将特征输入支持向量机进行识别。该模型在 H. sapiens 基准数据集上获得了 90.42%的总体准确率，但由于其未提供 Web 服务器，可用性低。为满足研究需求，邱等人基于随机森林算法建立了免费的网络服务器 iRNAm5C-PseDNC [14]，准确率为 92.37%，但数据集并未做相似筛选，导致模型高估。PEA-m5C [15]是另一种基于随机森林的算法，它

针对 *A. thaliana* 数据集的检验，具有高度不平衡的正/负比率，使其在排除误报的同时保持稳健，在 10 倍交叉验证中，总体准确率为 83.5%。RNAM5CPred [16]对三种类型的特征进行提取：KNFS (K-核苷酸频率)、pseDNC (伪二核苷酸组成)和 KSNPFs (K-间隔核苷酸对频率)，也由于对多种特征的提取，该模型的准确率为 92.5%。M5C 位点预测工具性能总结，如表 2 所示。

Table 2. Performance of the M5C modification site prediction tool
表 2. M5C 修饰位点预测工具的性能

Method	Species	ML-Algorithm	Sn	Sp	ACC	MCC
m5C-PseDNC	*H. sapiens*	SVM	0.850	0.958	0.904	0.810
iRNAm5C-PseDNC	*H. sapiens*	Random Forest	0.817	0.950	0.883	0.774
RNAm5CPred	*H. sapiens*	SVM	0.846	0.855	0.850	0.703
PEA-m5C	*A. thaliana*	Random Forest	0.432	0.454	0.443	−0.114

4.3. N1-甲基腺苷

目前，存在两个识别 N1 甲基腺苷位点的机器学习方法，即 RAMPred 和 ISGm1A。RAMPred [17]使用的特征编码方法是基于物理性质、化学性质和基本累积频率描述的特征的 41 nt 序列，采用 SVM 分类器对智人、肌肉支原体和酿酒酵母中的 m1A 修饰位点识别的模型。ISGm1A [18]是采用典型的序列特性，即核苷酸的物理和化学特性和累积频率，以及来自基因组注释的 75 个额外特性，基于随机森林算法的模型。Liu 等人通过对特征重要性的分析，发现了基因组特征在位点预测中的重要性，而大部分前人的算法研究大多忽略了基于注释的基因衍生的位点的拓扑信息。该模型在特征提取阶段，整合序列特征和基因组特征，获得了较好的结果。M1A 位点预测工具性能总结，如表 3 所示。

Table 3. Performance of the M1A modification site prediction tool
表 3. M1A 修饰位点预测工具的性能

Method	Species	ML-Algorithm	Sn	Sp	ACC	MCC
RAMPred	*M. musculus*	SVM	0.975	1.000	0.987	0.970
	H. sapiens		0.984	0.999	0.991	0.980
	S. cerevisiae		0.957	1.000	0.978	0.960
ISGm1A	*H. sapiens*	Random Forest	0.832	0.838	0.835	0.670

5. 结论

随着生物信息领域的发展 RNA 修饰在调节基因表达和疾病发病机制中的重要性，已被人们所熟知。近年来，对于 RNA 修饰位点的预测技术在理论深化和算法改进等方面都取得了一定的进展，但发展的过程中也发现了一些存在的问题。在论述研究的过程中，主要的发现是大多数 RNA 修饰位点是别的方法共享相同的技术、分类算法，但应用在相同或不同位点的识别表现结果均有所差异。其中，模型性能与基准数据集的质量和大小相关。除此之外，目前基于机器学习的预测模型的训练数据集样本长时间未更新，取样的 RNA 修饰位点数据不够完善导致泛化能力不强，且对于一些实验室的数据集没有明确的衡量基准，不同数据集训练模型的结果对于模型间的性能比较，有失偏颇。其次，所采用的分类算法大体还是以传统分类算法 SVM 为主，只有部分预测模型，采用了深度学习中的卷积神经网络 CNN。再者，从预

测结果上看，还有一定的提升空间。

未来的研究工作可围绕着所存在的已知问题开展，扩大数据集规模，建立明确的数据集衡量标准，增加物种数量，利用深度学习算法进一步提高 RNA 甲基化位点预测精度，为基因组学的研究打下基础。由人工神经网络发展而来的深度学习，其算法模型拥有更强的泛化能力，对未知数据集有更准确的拟合结果，大量数据集的训练下的深度学习算法，可提高 RNA 甲基化修饰位点的预测准确率。基于深度学习模型的 RNA 甲基化修饰位点的预测将是未来的研究方向之一。

参考文献

[1] Meyer, K.D. and Jaffrey, S.R. (2014) The Dynamic Epitranscriptome: N6-methyladenosine and Gene Expression Control. *Nature Reviews Molecular Cell Biology*, **15**, 313-326. https://doi.org/10.1038/nrm3785

[2] Djebali, S., Davis, C.A., Merkel, A., *et al*. (2012) Landscape of Transcription in Human Cells. *Nature*, **489**, 101-108. https://doi.org/10.1038/nature11233

[3] Hauenschild, R., Tserovski, L., Schmid, K., Thüring, K., Winz, M.L., Sharma, S., Entian, K.D., Wacheul, L., Lafontaine, D.L. anderson, J., Alfonzo, J., Hildebrandt, A., Jäschke, A., Motorin, Y. and Helm, M. (2015) The Reverse Transcription Signature of N-1-methyladenosine in RNA-Seq Is Sequence Dependent. *Nucleic Acids Research*, **43**, 9950-9964. https://doi.org/10.1093/nar/gkv895

[4] Bohnsack, K.E., Höbartner, C. and Bohnsack, M.T. (2019) Eukaryotic 5-methylcytosine (m5C) RNA Methyltransferases: Mechanisms, Cellular Functions, and Links to Disease. *Genes*, **10**, 102. https://doi.org/10.3390/genes10020102

[5] Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S. and Soboleva, A. (2013) NCBIGEO: Archive for Functional Genomics Data Sets—Update. *Nucleic Acids Research*, **41**, D991-D995. https://doi.org/10.1093/nar/gks1193

[6] Xuan, J.J., Sun, W.J., Lin, P.H., Zhou, K.R., Liu, S., Zheng, L.L., Qu, L.H. and Yang, J.H. (2018) RMBase v2.0: Deciphering the Map of RNA Modifications from Epitranscriptome Sequencing Data. *Nucleic Acids Research*, **46**, D327-D334. https://doi.org/10.1093/nar/gkx934

[7] Chen, W., Feng, P., Ding, H., Lin, H. and Chou, K.C. (2015) iRNA-Methyl: Identifying n6-methyladenosine Sites Using Pseudo Nucleotide Composition. *Analytical Biochemistry*, **490**, 26-33. https://doi.org/10.1016/j.ab.2015.08.021

[8] Zhou, Y., Zeng, P., Li, Y.H., *et al*. (2016) SRAMP: Prediction of Mammalian N6-methyladenosine (m6A) Sites Based on Sequence-Derived Features. *Nucleic Acids Research*, **44**, e91. https://doi.org/10.1093/nar/gkw104

[9] Qiang, X., Chen, H., Ye, X., *et al*. (2018) M6AMRFS: Robust Prediction of N6-methyladenosine Sites with Sequence-Based Features in Multiple Species. *Frontiers in Genetics*, **9**, Article No. 495. https://doi.org/10.3389/fgene.2018.00495

[10] Nazari, I., Tahir, M., Tayara, H., *et al*. (2019) iN6-Methyl (5-Step): Identifying RNA N6-methyladenosine Sites Using Deep Learning Mode via Chou's 5-Step Rules and Chou's General PseKNC. *Chemometrics and Intelligent Laboratory Systems*, **193**, Article ID: 103811. https://doi.org/10.1016/j.chemolab.2019.103811

[11] Alam, W., Ali, S.D., Tayara, H., *et al*. (2020) A CNN-Based RNA N6-methyladenosine Site Predictor for Multiple Species Using Heterogeneous Features Representation. *IEEE Access*, **8**, 138203-138209. https://doi.org/10.1109/ACCESS.2020.3002995

[12] Rehman, M.U., Hong, K.J., Tayara, H., *et al*. (2021) m6A-NeuralTool: Convolution Neural Tool for RNA N6-Methyladenosine Site Identification in Different Species. *IEEE Access*, **9**, 17779-17786. https://doi.org/10.1109/ACCESS.2021.3054361

[13] Feng, P., Ding, H., Chen, W., *et al*. (2016) Identifying RNA 5-methylcytosine Sites via Pseudo Nucleotide Compositions. *Molecular BioSystems*, **12**, 3307-3311. https://doi.org/10.1039/C6MB00471G

[14] Qiu, W.R., Jiang, S.Y., Xu, Z.C., *et al*. (2017) iRNAm5C-PseDNC: Identifying RNA 5-methylcytosine Sites by Incorporating Physical-Chemical Properties into Pseudo Dinucleotide Composition. *Oncotarget*, **8**, 41178-41188. https://doi.org/10.18632/oncotarget.17104

[15] Chen, Z., Zhao, P., Li, F., *et al*. (2020) iLearn: An Integrated Platform and Meta-Learner for Feature Engineering, Machine-Learning Analysis and Modeling of DNA, RNA and Protein Sequence Data. *Briefings in Bioinformatics*, **21**, 1047-1057. https://doi.org/10.1093/bib/bbz041

[16] Fang, T., Zhang, Z., Sun, R., *et al*. (2019) RNAm5CPred: Prediction of RNA 5-methylcytosine Sites Based on Three Different Kinds of Nucleotide Composition. *Molecular Therapy-Nucleic Acids*, **18**, 739-747. https://doi.org/10.1016/j.omtn.2019.10.008

[17] Chen, W., Feng, P., Tang, H., *et al*. (2016) RAMPred: Identifying the N1-methyladenosine Sites in Eukaryotic Transcriptomes. *Scientific Reports*, **6**, Article No. 31080. https://doi.org/10.1038/srep31080

[18] Liu, L., Lei, X., Meng, J., *et al*. (2020) ISGm1A: Integration of Sequence Features and Genomic Features to Improve the Prediction of Human m1A RNA Methylation Sites. *IEEE Access*, **8**, 81971-81977. https://doi.org/10.1109/ACCESS.2020.2991070