Hans Journal of Biomedicine 生物医学, 2022, 12(2), 109-115
Published Online April 2022 in Hans. http://www.hanspub.org/journal/hjbm
https://doi.org/10.12677/hjbm.2022.122014

Hans汉斯

基于机器学习的假尿苷位点预测的研究进展

孟　蕊

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3月17日；录用日期：2022年4月8日；发布日期：2022年4月15日

摘　要

在基因的转录过程中，RNA很容易发生修饰的现象。迄今为止，研究人员已经发现了一百多种RNA的修饰，而假尿苷(ψ)是第一个被发现的，并且是目前存在最广泛的一种RNA修饰。近年来，随着表观遗传学研究的深入，关于假尿苷的研究越来越多。假尿苷修饰对于各种细胞生物和生理过程是至关重要的，研究的关键步骤就是在转录组中准确地识别出假尿苷的位点。由于实验化学方法来识别假尿苷位点耗时耗力，基于机器学习的计算方法来识别假尿苷位点是如今最好的选择。本文回顾了基于机器学习的假尿苷位点预测的研究现状，调查了研究人员在位点预测过程中使用的数据集和评估方法，得到了假尿苷位点预测的最新进展。本文选取具有代表性的几个机器学习模型进行简要概述，并对目前的局限性给出一些建议。

关键词

RNA修饰，假尿苷，位点预测，机器学习

Research Progress of Pseudouridine Site Prediction Based on Machine Learning

Rui Meng

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

Received: Mar. 17th, 2022; accepted: Apr. 8th, 2022; published: Apr. 15th, 2022

Abstract

RNA is easily modified in the process of gene transcription. To date, researchers have found more than a hundred RNA modifications, and pseudouridine (ψ) was the first to be discovered and is the most widely available type of RNA modification. In recent years, with the development of epigenetics, more and more studies on pseudouridine have been conducted. Pseudouridine modification is essential for various cellular biological and physiological processes, and the key step of the

文章引用: 孟蕊. 基于机器学习的假尿苷位点预测的研究进展[J]. 生物医学, 2022, 12(2): 109-115.
DOI: 10.12677/hjbm.2022.122014

study is to accurately identify pseudouridine sites in the transcriptome. Since it is time-consuming and labor-intensive to identify pseudouridine sites by experimental chemical methods, the computational method based on machine learning is the best choice to identify pseudouridine sites today. This paper reviews the research status of pseudouridine site prediction based on machine learning, investigates the data sets and evaluation methods used by researchers in the process of site prediction, and gets the latest progress on pseudouridine site prediction. In this paper, several representative machine learning models are selected to give a brief overview and provide some suggestions on the current limitations.

Keywords

RNA Modification, Pseudouridine, Site Prediction, Machine Learning

1. 介绍

假尿苷(Ψ)是一种必不可少且普遍存在的 RNA 修饰类型，被称为“第五种 RNA 核苷酸”，在真核生物和原核生物的多种类型的 RNA 中广泛发现了这种修饰，包括 tRNA，mRNA 和 rRNA [1]。大量研究表明，假尿苷在稳定 RNA 结构[2] [3]，RNA-蛋白质或 RNA-RNA 相互作用[4]，调节进入位点结合过程[5]以及 RNA 的代谢[6] [7]等分子机制中起着至关重要的作用。因此假尿苷位点的识别对于揭示相关的生物学原理至关重要。

假尿苷是尿苷的同分异构体，在 RNA 中假尿苷的形成主要有两种机制。一种是由高度保守的蛋白质也就是假尿苷合酶催化的。这个假尿苷合酶同时起到识别和催化的两种作用。即将尿苷残基的碱基从糖中分离出来，将其沿着 N3-C6 轴“旋转”180 度，然后将碱基的 5-碳重新连接到糖的 1-碳上。假尿苷修饰位点结构如图 1 所示。另外一种是依赖于一类 snRNA 与相应的蛋白质形成的复合物，RNA 起到识别作用，与其结合的蛋白质发挥催化作用[8]。

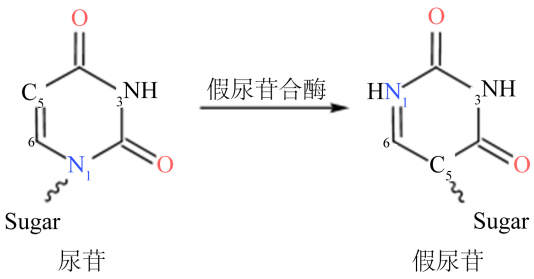

Figure 1. Pseudouridine modified site structure
图 1. 假尿苷修饰位点结构

尽管 RNA 假尿苷修饰在几十年前就被发现了，但随着下一代测序技术的迅速发展，第一个全转录组 RNA 假尿苷修饰图谱直到 2014 年才发表。Carlile [7]等人开发了 PseudoU-seq 技术，他们利用该技术在酵母和人类细胞的受调控 mRNA 中鉴定了 200 多个假尿苷化位点，同年，Schwartz [9]等人利用类似的方

法进行了全转录组作图，在非编码 RNA 和 mRNA 中发现了 300 多个动态调控的假尿苷位点。Li [10]等人提出了一种化学标记方法(CeU-Seq)，他们在人类 mRNA 中标记了超过 2000 个假尿苷位点。其他的研究人员也开发了其他 RNA 假尿苷测序的方案。虽然这些实验方法和化学方法在预测假尿苷位点的过程中发挥着重要作用，但是工作量比较大，不仅昂贵且花费的时间和精力也很多。由于后基因组时代产生的数据量不断增加，最近出现了用于 RNA 化学修饰预测的稳健、快速和廉价的计算方法，多数基于传统的机器学习算法，也有一些基于深度学习算法。在本文中，对近年来基于机器学习的假尿苷位点预测的模型进行介绍。

2. 通用数据集

在 2016 年，Chen [11]等人基于 RMBase 建立了第一个基准数据集，分别命名为 H_990 (人类)、S_628 (酿酒酵母)和 M_944 (小家鼠)用于模型训练；以及另外两个独立的测试数据集，名为 H_200 (人类)和 S_200 (酿酒酵母)，用于不同方法之间的性能验证和比较。在 2019 年，Liu [12]等人基于 RMBase v2.0 更新了训练数据集，并获得三个新的训练数据集 NH_990 (人类)、NM_944 (小家鼠)和 NS_627 (酿酒酵母)，分别比原始数据集中的人类，酿酒酵母和小家鼠多 26、10 和 1 个样本。由于这两个常用且数据集差别很小，且本文介绍的预测模型都是采用 Chen 等人的数据集，因此对 Chen 等人建立的数据集进行说明。基准数据集中人类训练数据集包含 495 个假尿苷位点序列和 495 个非假尿苷位点序列；酿酒酵母训练数据集包含 314 个假尿苷位点序列和 314 个非假尿苷位点序列；小家鼠训练数据集包含 944 个序列，其中一半为阳性样本。测试数据集人类和酿酒酵母均含有 100 个阳性样品和 100 个阴性样品。人类和小家鼠数据集中的 RNA 序列均包含 21 个核苷酸，酿酒酵母数据集中的 RNA 序列包含 31 个核苷酸。数据集如表 1 所示。

Table 1. Base dataset
表 1. 基准数据集

物种	交叉验证		独立测试		长度(bp)
人类	900	495 阳性 495 阴性	200	100 阳性 100 阴性	21
酿酒酵母	628	314 阳性 314 阴性	200	100 阳性 100 阴性	31
小家鼠	944	472 阳性 472 阴性	/	/	21

3. 模型性能评估标准

在生物信息学和最近的研究领域，四个指标被用来评价预测因子的质量。它们是特异性(SP)，敏感性(SN)，准确性(ACC)和马修斯相关系数(MCC)。公式如下：

$$SN = \frac{TP}{TP + FN}$$

$$SP = \frac{TN}{TN + FP}$$

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

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

其中 TP (真阳性)的含义是本身是假尿苷位点，也被预测为假尿苷位点；TN (真阴性)为本身是非假尿苷位点，也被预测为非假尿苷位点；FN (假阴性)为本身为假尿苷位点，但是被预测为非假尿苷位点；FP (假阳性)为本身为非假尿苷位点，但是被预测为假尿苷位点。因此，SN 是准确预测假尿苷位点的可能性。SP 是获得非假尿苷位点的正确预测的可能性。ACC 代表整体 RNA 序列位点预测的准确性。由于 MCC 考虑到真阳性，假阳性，真阴性和假阴性四个特征，它通常被视为衡量平衡的尺度。

4. 预测模型

在 2020 年，Lv [13]等人，提出了一种名为 RF-PseU 的随机森林预测器用于预测假尿苷的位点。随机森林算法是一种袋式集成学习算法[14]。通过组合多个弱分类器，最终的结果可以投票或平均，以获得一个更高的精度，更好的综合性能和抗过拟合的整体模型。这个算法已经被广泛使用在生物信息学和其他领域的应用，并在各个领域已被证实是一种有效的建模技术。

为了确定最优的特征空间，Lv 等人首先使用梯度增强算法(LGBM)根据特征的重要性值将特征从最大值排序到最小值。所有重要性值大于平均值的特征都要被保留。其次，使用了增量特征选择策略(IFS)，随着特征的添加，交叉验证和独立测试的精度都发生了变化，起初每个物种的准确性增加的比较迅速，后呈波形平缓。便于比较，文章使用了 LOO 交叉验证和 10 倍交叉验证两种交叉验证方法，得到两个结果来评估训练模型。RF-PseU 的优势在同水平下，开发了一个具有易于使用的界面的 web 服务器，以便于相关用户使用和研究。

在 2021 年，Li [15]等人提出了一个名叫 Porpoise 的堆叠集成机器学习框架，旨在改进 RNA 假尿苷位点的预测。Porpoise 对 18 种特征编码方案和 9 种常用机器学习算法的性能进行了全面的基准测试，对于每种机器学习算法，根据每种特征类型训练 18 个分类器，并根据马修斯相关系数(MCC)选择性能最好的一个作为候选基分类器。使用 Python 中的 scikit-learn 包，通过 10 次 10 倍的交叉验证测试，构建和优化了所有分类器。使用这种策略，根据 9 种不同的机器学习算法共获得了 9 个候选基分类器。由于 9 个基分类器的预测性没有得到满意的结果，因此采用叠加策略建立了集成学习模型。叠加是一种有效的集成学习策略，它综合了各种分类器的信息，从而能够建立一个稳健的预测模型。这种策略已经成功应用于最近的一些生物信息学和计算生物学的研究。堆叠策略包含两个主要步骤，每个步骤中相应的分类器被称为基分类器和元分类器。第一步应用并建立一组基本分类器，第二步以基本分类器的输出作为输入，对元分类器进行训练。文章中首先根据 9 个基本分类器的分类性能对其进行了排序，设置 c 为候选基分类器的排序池，c = {c1, c2, c3, c4, c5, c6, c7, c8, c9}，其中 c1 获得了最好的 MCC，从 c 获取基本分类器生成八个基本分类器组合，其中 Ensemble1 包括{c1, c2}，Ensemble 2 包括{c1, c2, c3}，以此类推，直到 Ensemble 8 包括{c1, c2, ⋯, c8, c9}。使用 logistic 回归作为元分类器来训练堆积模型，最终选择获得最佳性能的组合作为最终模型。

在基本分类器优化组合的基础上，Porpoise 对每个基本分类器进行特征选择和超参数优化。采用了两步特征选择策略，应用了 mRMR 特征选择算法(最小冗余最大相关性)来对特征进行排序，并使用增量特征选择算法(IFS)来选择最佳特征。选择达到最高 MCC 值的特征子集为最佳特征子集。之后采用贝叶斯优化算法来优化堆叠模型的超参数。此外，文章中还确认了所选特征的重要性，并使用 SHAP 算法帮助解释 Porpoise 的堆叠模型。

在 2021 年，Wang [16]等人提出了一种新的特征融合预测器，命名为 PsoEL-PseU，用于预测假尿苷位点。首先，本研究系统全面地探索了不同类型的特征编码方案，确定了具有不同性质的六种特征编码方案。为了提高特征表示能力，充分利用这些特征编码方案，使用二进制粒子群优化机器学习算法来消除大量冗余和无效的特征，从而捕获六个特征编码方案的最优特征子集。其次，六个个体预测器通过使

用六个最佳特征子集进行训练。最后，为了融合所有六个特征的效果，通过并行融合策略将六个个体预测器融合到集合预测器中。其中并行融合策略采取的是多数表决策略。在三个基准数据集上的十倍交叉验证表明 PsoEL-PseU 预测器性能得到了明显提升。PsoEL-PseU 也提供了一个用户友好型网络服务器可以自由访问。

在 PsoEL-PseU 的粒子群优化算法中，每个粒子由两部分组成，其中第一部分代表特征选择的结果。它的长度等于原始特征的数量。粒子位置的 0 或 1 的值用来表示是否选择了相应位置的特征。第二部分是由 10 个二进制位组成的支持向量机(SVM)训练的超参数组合的结果，它可以表示总共 1024 个超参数组合。最后，为了获得更有效的特征，将已识别的伪尿苷位点的十倍交叉验证的分类精度作为适应值，以保证种群粒子向高分类精度的方向移动。详细的流程图如图 2。首先初始化粒子，然后计算每个粒子的适应度值。适应度值用于迭代更新速度和位置，以找到特征描述符的最优特征子集。

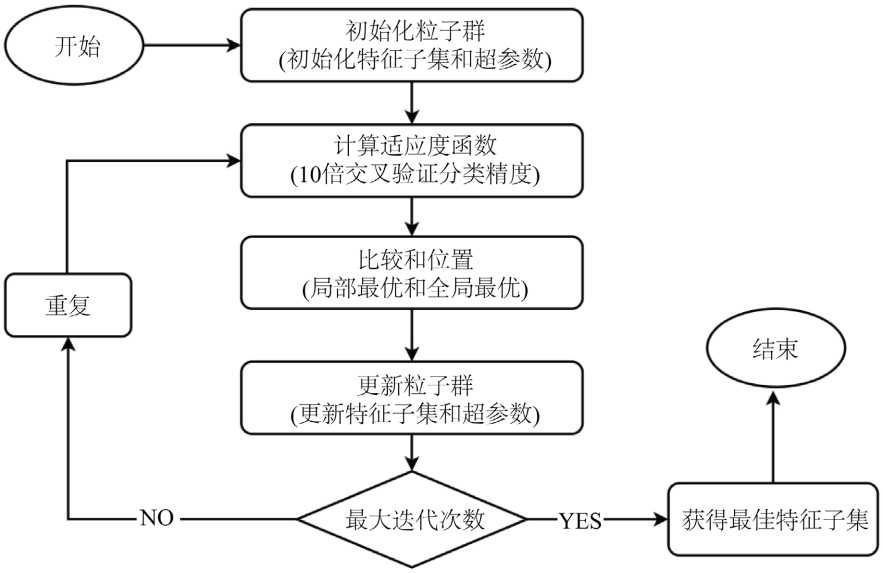

Figure 2. Flow chart of the BPSO algorithm
图 2. BPSO 算法流程图

5. 结果比较

表 2 为三种机器学习模型预测假尿苷位点在训练集上的表现比较。三种模型在三个物种中 ACC 都达到了 60%以上，能准确地预测假尿苷位点，且在酿酒酵母和小鼠中准确度更高。在准确预测位点的可能性上，Porpoise 表现最好，比其他两种模型高出 22.21%，3.01%，4.73%以上。Porpoise 的 MCC 分数人类和酵母细菌中达到了最高，但是在小家鼠中，PsoEL-PseU 的 MCC 分数最好。

Table 2. Comparison of training set performance
表 2. 训练集表现比较

物种	模型	ACC (%)	Sn (%)	Sp (%)	MCC
人类 (H_990)	Porpoise	**78.53**	**89.11**	67.94	**0.585**
	PsoEL-PseU	70.8	66.9	**74.7**	0.42
	RF-PseU(10 Fold)	64.3	66.1	62.6	0.29
	RF-PseU (LOO)	64	65.9	62.6	0.29

Continued

酿酒酵母(S_628)	Porpoise	**81.69**	**81.21**	82.17	**0.634**
	PsoEL-PseU	80.3	69.1	**91.4**	0.62
	RF-PseU(10 Fold)	74.8	77.2	72.4	0.49
	RF-PseU (LOO)	75.8	78.2	73.4	0.52
小家鼠(M_944)	Porpoise	**77.75**	**77.83**	77.67	0.556
	PsoEL-PseU	76.5	53	**82.2**	**0.708**
	RF-PseU(10 Fold)	74.8	73.1	76.5	0.5
	RF-PseU (LOO)	74.5	72.7	75.2	0.48

表 3 为三种机器学习模型预测假尿苷位点在独立测试基上的表现比较。在测试集上，除了 Sp 的值，Porpoise 的得分都是最好的，在 H_200 上的准确性为 77.35%，MCC 为 0.551，与训练数据集的结果非常接近。而在 S_200 上的准确性为 83.5%，MCC 为 0.673，略高于训练数据集的准确率。

Table 3. Comparison of test set performance
表 3. 测试集表现比较

物种	模型	ACC (%)	Sn (%)	Sp (%)	MCC
人类 (H_200)	Porpoise	**77.35**	**82.3**	72.4	**0.551**
	PsoEL-PseU	75.5	76	**75**	0.51
	RF-PseU (10 Fold)	75	78	72	0.5
	RF-PseU (LOO)	74	74	74	0.48
酿酒酵母 (S_200)	Porpoise	**83.5**	**88**	79	**0.673**
	PsoEL-PseU	82	83	**81**	0.64
	RF-PseU (10 Fold)	77	75	79	0.54
	RF-PseU (LOO)	74.5	70	79	0.49

6. 结论与展望

RNA 修饰的研究引起了人们的高度关注，因为它揭示了 RNA 修饰在调节基因表达和疾病发病机制中的重要性。随着表观转录组测序数据的增加，更多的 RNA 修饰基准数据集变得可用。最近大型数据集的可用性以及通过机器学习在计算生物学方面取得的进步已经改变了该领域的研究。因此，这些技术最终提高了我们对 RNA 修饰生物学意义的理解。而对于 RNA 修饰中最常见的假尿苷修饰，其在结构功能和新陈代谢中起到了重要的作用，因此准确的识别假尿苷的位点对揭示有关生物学原理至关重要。而由于实验化学方法的费时费力，开发出基于机器学习的计算方法来预测假尿苷位点是很有必要的。

本文回顾了基于机器学习的假尿苷位点预测的最新进展，尽管预测算法一直在更新迭代，但是目前的模型仍然有一些限制和问题。为了打破局限性，获得更好的预测精度，本文给出两点改进和提高的建议：

1) 机器学习在学术研究领域和实际应用领域得到越来越多的关注，并且展示了其独特的优势。但是在生物序列如 DNA 序列，RNA 序列等方面的应用还比较少，后续工作可以深入研究学习算法的内部构成，做到可以改进内部结构以适应生物序列方面的数据集；或者可以开发出匹配生物序列的学习算法。

2) 多种 RNA 都含有假尿苷修饰，包括 tRNA，rRNA，mRNA。在预测假尿苷位点时根据不同类型的 RNA 进行位点预测，即在预测之前先判断是哪一种 RNA，不同的 RNA 有不同的预测分类器，从而使预测的精度得到提高。

参考文献

[1] Ge, J. and Yu, Y.-T. (2013) RNA Pseudouridylation: New Insights into an Old Modification. *Trends in Biochemical Sciences*, **38**, 210-218. https://doi.org/10.1016/j.tibs.2013.01.002

[2] Charette, M. and Gray, M.W. (2000) Pseudouridine in RNA: What, Where, How, and Why. *IUBMB Life*, **49**, 341-352. https://doi.org/10.1080/152165400410182

[3] Davis, D.R., Veltri, C.A. and Nielsen, L. (1998) An RNA Model System for Investigation of Pseudouridine Stabilization of the Codonanticodon Interaction in $tRNA^{Lys}$, $tRNA^{His}$ and $tRNA^{Tyr}$. *Journal of Biomolecular Structure and Dynamics*, **15**, 1121-1132. https://doi.org/10.1080/07391102.1998.10509006

[4] Basak, A. and Query, C.C. (2014) A Pseudouridine Residue in the Spliceosome Core Is Part of the Filamentous Growth Program in Yeast. *Cell Reports*, **8**, 966-973. https://doi.org/10.1016/j.celrep.2014.07.004

[5] Jack, K., Bellodi, C., Landry, D.M., Niederer, R.O., Meskauskas, A., Musalgaonkar, S., *et al*. (2011) rRNA Pseudouridylation Defects Affect Ribosomal Ligand Binding and Translational Fidelity from Yeast to Human Cells. *Molecular Cell*, **44**, 660-666. https://doi.org/10.1016/j.molcel.2011.09.017

[6] Ma, X., Zhao, X. and Yu, Y.T. (2003) Pseudouridylation (Ψ) of U2 snRNA in *S. cerevisiae* Is Catalyzed by an RNA-Independent Mechanism. *The EMBO Journal*, **22**, 1889-1897. https://doi.org/10.1093/emboj/cdg191

[7] Carlile, T.M., Rojas-Duran, M.F., Zinshteyn, B., Shin, H., Bartoli, K.M., Gilbert, W.V., *et al*. (2014) Pseudouridine Profiling Reveals Regulated mRNA Pseudouridylation in Yeast and Human Cells. *Nature*, **515**, 143-146. https://doi.org/10.1038/nature13802

[8] 毕月. 基于机器学习的 RNA 相关功能位点研究[D]: [硕士学位论文]. 大连: 大连海事大学, 2020. https://doi.org/10.26989/d.cnki.gdlhu.2020.001025

[9] Schwartz, S., Bernstein, D.A., Mumbach, M.R., Jovanovic, M., Herbst, R.H. and León-Ricardo, B.X. (2014) Transcriptome-Wide Mapping Reveals Widespread Dynamic-Regulated Pseudouridylation of ncRNA and mRNA. *Cell*, **159**, 148-162. https://doi.org/10.1016/j.cell.2014.08.028

[10] Li, X.Y., Zhu, P., Ma, S.Q., Song, J.H., Bai, J.Y., Sun, F.F., *et al*. (2015) Chemicalpulldown Reveals Dynamic Pseudouridylation of the Mammalian Transcriptome. *Nature Chemical Biology*, **11**, 592-597. https://doi.org/10.1038/nchembio.1836

[11] Chen, W., Tang, H., Ye, J., Lin, H. and Chou, K.C. (2016) iRNA-PseU: Identifying RNA Pseudouridine Sites. *Molecular Therapy*: *Nucleic Acids*, **5**, Article ID: e332

[12] Liu, K., Chen, W. and Lin, H. (2020) XG-PseU: An Extreme Gradient Boosting-Based Method for Identifying Pseudouridine Sites. *Molecular Genetics and Genomics*, **295**, 13-21. https://doi.org/10.1007/s00438-019-01600-9

[13] Lv, Z., Zhang, J., Ding, H. and Zou, Q. (2020) RF-PseU: A Random Forest Predictor for RNA Pseudouridine Sites. *Frontiers in Bioengineering and Biotechnology*, **8**, Article No. 134. https://doi.org/10.3389/fbioe.2020.00134 https://www.frontiersin.org/article/10.3389/fbioe.2020.00134.

[14] Cheng, L., Hu, Y., Sun, J., Zhou, M. and Jiang, Q. (2018) DincRNA: A Comprehensive Web-Based Bioinformatics Toolkit for Exploring Disease Associations and ncRNA Function. *Bioinformatics*, **34**, 1953-1956. https://doi.org/10.1093/bioinformatics/bty002

[15] Li, F., Guo, X., Jin, P., Chen, J., Xiang, D., Song, J. and Coin, L.J.M. (2021) Porpoise: A New Approach for Accurate Prediction of RNA Pseudouridine Sites, *Briefings in Bioinformatics*, **22**, Article No. bbab245. https://doi.org/10.1093/bib/bbab245

[16] Wang, X., Lin, X., Wang, R., Han, N., Fan, K., Han, L. and Ding, Z. (2021) A Feature Fusion Predictor for RNA Pseudouridine Sites with Particle Swarm Optimizer Based Feature Selection and Ensemble Learning Approach. *Current Issues in Molecular Biology*, **43**, 1844-1858. https://doi.org/10.3390/cimb43030129