

引用格式:刘瑞,张晶,米浩源,等. 基于 SVR 模型的血液 DNA 甲基化年龄预测研究[J]. 北京化工大学学报(自然科学版), 2026, 53(4): 72-78.

LIU Rui, ZHANG Jing, MI HaoYuan, et al. A blood DNA methylation age - prediction model based on support vector regression[J]. Journal of Beijing University of Chemical Technology (Natural Science), 2026, 53(4): 72-78.

基于 SVR 模型的血液 DNA 甲基化年龄预测研究

刘 瑞 张 晶 米浩源 樊 埔
(中国人民公安大学 侦查学院, 北京 100038)

摘 要: 基于 Illumina 无限甲基化 EPIC 珠芯片(850K)检测的血液 DNA 甲基化数据,采用少维度、强线性位点相关位点筛选策略,利用支持向量回归(SVR)原理,建立高稳定性和准确度的血液脱氧核糖核酸(DNA)甲基化(methylation)年龄推断模型。通过文献调研进行 CpG 位点选择,利用 SVR 进行模型构建和年龄推断。采用 Matlab 进行模型搭建,对标准参数进行选取后,使用网格搜索(grid search)结合交叉验证(cross validation)的方法,对 SVR 模型中的惩罚因子 C 和核函数参数 g 进行系统优化,以确定模型性能最优的参数组合。数据处理时先对甲基化水平 β 进行归一化并代入模型,再通过反归一化得出预测年龄。所构建的模型在血液样本年龄预测中的决定系数(R^2)为 0.885 29,测试集平均绝对偏差(MAD)为 2.57 岁,表明模型整体预测精度较高;虽然预测误差随着年龄增长略有增大,但总体仍可接受。模型预测准确度随着年龄的增长而下降。本研究验证了仅在使用 4 个 CpG 位点的情况下,SVR 模型依然能够实现高精度的年龄预测,预测误差随着年龄增长略有增加但总体稳定,表明该模型具有良好的泛化能力和实用价值。该方法可为法医学中犯罪嫌疑人生理特征刻画提供有效技术支持。

关键词: 法医遗传学; 支持向量回归(SVR); 年龄推断; DNA 甲基化(DNA_m)

中图分类号: R89; Q523 **DOI:** 10.13543/j.bhxbzr.2026.04.008

引 言

脱氧核糖核酸(deoxyribonucleic acid, DNA)甲基化(methylation)年龄推断方法具有广泛的应用前景和科学意义。DNA 甲基化(DNA_m)模式在人类整个生命周期中随年龄增长呈现动态且规律性的变化。通过比较个体的实际年龄与其基因组中的 DNA_m 水平,可以建立预测模型来准确推断生理年龄。DNA_m 可用于评估生理和健康状况,探索其与衰老、疾病风险之间的关系,并为个性化医疗和健康管理提供参考^[1-3]。通过比较 DNA_m 年龄和实足年龄,可以及早识别潜在的健康风险,从而提前采取适当的预防措施^[4-5]。此外,分析不同群体间 DNA_m 年龄的差异,可以揭示环境因素(如生活方式、饮食和污染等)对生理年龄的影响,为制定卫生

政策和干预措施提供依据^[6-7]。

在法医实践中, DNA_m 年龄预测为确定死亡原因和推断死后间隔时间提供了新的方法。通过分析 DNA_m 水平来推断生理年龄,可为法律实践提供关键的信息。首先,通过对提取样本的 DNA_m 分析可准确推断嫌疑人的年龄,从而辅助刑事调查;其次, DNA_m 年龄推断有助于在涉及身份不明的个体案件中确定涉案人员的年龄,例如尸体检查或儿童诱拐案件;最后, DNA_m 年龄推断还有助于判定刑事责任年龄,因为在许多法律体系中,未成年人和成年人的量刑年龄有所不同^[8]。在某些案件中,现场仅遗留血液、精液等生物物证,无法通过短串联重复序列(short tandem repeat, STR)数据库比对来获取有效身份信息。若能利用甲基化数据推断嫌疑人大致年龄,则可显著缩小排查范围,提高侦查效率。同时,该方法还可作为身份认定的辅助验证手段,与其他生理特征(如性别、种族、外貌预测等)共同构建更为精准的嫌疑人画像。

随着表观遗传学领域的发展,人类特征研究的重点已从限制性片段长度多态性(RFLP)和 STR 等传统遗传标记,扩展成单核苷酸多态性(SNPs)和

收稿日期:2024-12-10

基金项目:中国人民公安大学刑事科学技术双一流创新研究专项(2023SYL06)

第一作者:女,1987年生,副教授

E-mail:liurui1987110@126.com

DNA_m等表观遗传标记^[9-10]。DNA_m作为一种重要的表观遗传标记,是对上述传统标记的补充,在基于其全生命周期的动态变化的年龄推断模型中发挥着关键作用^[4,11]。传统的甲基化模型通常采用线性回归,所需标记物和样本数量较多,增加了复杂性和成本^[12]。血液DNA_m年龄推断一直是一个活跃的研究领域。以往的研究利用焦磷酸测序和Epi-TYPER系统等技术,基于不同胞嘧啶-磷酸-鸟嘌呤(cytosine phosphate guanine, CpG)位点建立年龄预测模型,报告了不同水平的平均绝对偏差(mean absolute deviation, MAD)^[13-14]。这些模型的发展伴随着支持向量回归(support vector regression, SVR)模型的改进,旨在提高基于生物样本的年龄预测的稳健性和适用性^[8,15]。然而,现有的SVR模型仍存在成本高、设备兼容性不足、准确度低和稳定性有待提升等问题。

在我国法医学快速发展背景下,开发一套基于国产微滴式数字聚合酶链式反应平台(droplet digital polymerase chain reaction, ddPCR)的甲基化年龄推断模型具有明确的实际意义和推广价值。本研究探索了采用SVR建模方法进行年龄预测,仅选用少量强相关CpG位点,建立高度稳定、准确的血液DNA_m年龄预测模型。在本研究中,采用样本数据对多元线性回归(MLR)模型和SVR模型进行了比较分析,结果表明SVR模型在建立基于血液DNA的准确年龄预测模型方面具有更高的重复性、更好的拟合性和更强的兼容性。本研究旨在从血液DNA中筛选常见的CpG位点,利用Illumina无限甲基化EPIC珠芯片(850K)检测其DNA_m水平,进而建立基于血液的DNA_m年龄预测的SVR回归模型。本研究不仅丰富了血液DNA_m年龄预测的理论体系,也为我国法医学实践提供了可操作的技术手段,为身份识别与案件侦查提供科学依据。

1 实验部分

1.1 实验数据来源

本研究使用由公安部鉴定中心采集并检测的91份中国个体血液样本的数据。所有样本均遵守当事人知情同意原则,并通过公安部鉴定中心伦理委员会伦理审查(审批编号:2021-030)。实验采用的血液样本年龄范围为20~59岁,平均年龄39.08岁,分布情况如表1所示。

1.2 实验仪器

血液样本数据基于Illumina无限甲基化EPIC珠

表1 血液样本分布信息

年龄阶段/岁	样本量/个	平均年龄/岁
20~29	30	23.23
30~39	22	34.73
40~49	17	44.47
50~59	22	53.82

芯片(850K)获得。该芯片是一种基于微阵列的高通量DNA甲基化检测平台,广泛应用于表观遗传学研究。其基本检测原理:样本DNA首先经亚硫酸氢盐处理,将未甲基化的胞嘧啶(C)转化为尿嘧啶(U),而甲基化的胞嘧啶保持不变;亚硫酸氢盐处理后的DNA片段经PCR扩增,与芯片上预先固定的探针杂交。芯片上含有超过850 000个针对全基因组CpG位点的探针,覆盖启动子区、增强子区、CpG岛、岸边等关键调控区域。每个位点设计有两种探针,分别对应甲基化(M)与非甲基化(U)状态,通过单碱基延伸反应和不同的荧光标记进行区分。最终甲基化水平以 β 值表示,其范围从0(完全未甲基化)到1(完全甲基化)。

芯片通过激光扫描仪读取荧光信号强度,结合内置标准化算法,输出每个位点的甲基化 β 值数据,供后续统计分析与建模使用。

1.3 实验方法

1.3.1 CpG位点的筛选

本研究通过系统检索PubMed和Web of Science数据库中2015年以来的文献,参考多项研究在血液中重复验证的CpG位点^[16-17],初步筛选出8个位点。筛选标准包括:(1)文献报道中与年龄高度相关,Pearson相关系数(r) $r > 0.75$;(2)在不同研究中重复出现;(3)甲基化水平随年龄呈现持续增加或减少的稳定趋势。将所选位点进行编号如表2所示。

1.3.2 Pearson相关性分析

Pearson相关性分析是评估2个变量线性相关程度及方向的方法,其相关系数可反映变量间的线性关系强度。本研究使用R语言中“cor()”函数,对经Illumina无限甲基化EPIC珠芯片(850K)检测的91个样本数据进行分析,计算8个CpG位点甲基化程度与样本年龄的 r ,并绘制回归散点图。

1.3.3 位点数据的选择与处理优化

基于SVR对输入变量的要求,本研究从 $|r| > 0.5$ 的CpG位点中,选取离散较少的4个位点用于后续的SVR模型训练与构建。在Matlab软件平台,使用

表2 CpG 位点编号信息
Table 2 Numbering of CpG sites

位点编号	cg 编号
CPG1	cg21572722
CPG2	cg19283806
CPG3	cg03607117
CPG4	cg13552692
CPG5	cg14692377
CPG6	cg25410668
CPG7	cg04875128
CPG8	cg04453050

“mapminmax(‘apply’)”函数对数据进行归一化处理,以消除不同量纲对模型的影响。在预测数值输出时,使用“mapminmax(‘reverse’)”函数进行反归一化处理,得到正常数量级下的年龄预测数值。

1.3.4 年龄推断模型的构建

将筛选后的样本打乱,随机抽取 60 个样本作为模型的训练集,31 个样本作为模型的测试集。基于 Matlab 软件平台和 libsvm-3.34 工具包,使用网格搜索(grid search)与交叉验证(cross-validation)寻找最佳惩罚因子 C 和核函数参数 γ ,并带入“svmtrain()”函数,建立训练集样本年龄与位点甲基化水平间的 SVR 预测模型。采用“svmpredict()”函数对测试集样本进行年龄预测,并通过绘图与相关误差指标计算对模型进行可视化与误差分析。本研究采用 5 折交叉验证法(5-fold cross-validation)对年龄推断模型进行训练。

1.3.5 推断模型的准确性分析

将筛选后的样本分为 4 个不同年龄段:20~29 岁、30~39 岁、40~49 岁、50~59 岁,分别评估模型在各年龄组的预测性能,计算各组预测指标和准确率。若预测偏差在 ± 5 岁范围内,则认为推断结果准确可靠。

2 结果与讨论

2.1 Pearson 相关性

基于 Illumina 无限甲基化 EPIC 珠芯片(850K)检测 91 份血液样本中 8 个 CpG 位点的 DNA 甲基化水平,其与样本年龄的回归散点图如图 1 所示。计算所得 r 见表 3。

2.2 位点选择与数据处理

基于 Pearson 相关性分析结果,在 5 个 $|r|$ 值大于

0.5 的候选位点(CpG2、CpG3、CpG4、CpG5、CpG8)中,CpG8 的离散点较多,因此本研究最终选择 CpG2、CpG3、CpG4、CpG5 共 4 个位点建立模型。筛选后的样本数据在 Matlab 软件平台中使用“mapminmax(‘apply’)”函数进行归一化处理,将年龄数据和甲基化水平数据映射至(0,1)的区间内。

2.3 血液 DNA 甲基化年龄推断模型的构建

在建立 DNA 甲基化年龄预测模型的过程中,建模方法的选择直接关系到模型的拟合能力和泛化性能。尽管多元线性回归(multiple linear regression, MLR)具有方法简便、可解释性强等优点,在前期研究中被广泛采用,但该方法是基于自变量与因变量间存在线性关系的假设,难以有效捕捉复杂的非线性特征。

CpG 位点甲基化水平与年龄之间的关系在不同年龄段或不同基因区域常表现出非线性甚至平台期变化的特点,简单的线性模型可能导致拟合不足或偏差累积。为克服这一问题,本研究选用 SVR 算法。SVR 模型具备良好的非线性建模能力,能够通过引入核函数(如径向基核等),将输入特征映射到高维空间,在高维特征空间中寻找最优拟合超平面,从而更准确地刻画 CpG 甲基化水平随年龄变化的复杂趋势。

此外,SVR 模型在小样本、高维数据场景下具有良好的鲁棒性和泛化能力,这一特性使其特别适用于本研究中样本数量有限、特征位点数较少的情况,可有效平衡拟合精度与模型复杂度,避免过拟合现象。

将 91 份样本按约 2:1 的比例随机划分为训练集 60 份(约 66%)和测试集 31 份(约 34%)。基于 CpG2-CpG3-CpG4-CpG5 位点组合构建年龄推断模型。构建年龄推测模型的数学公式如下。

$$\mathbf{F}\{\mathbf{x}\} = [x_1, x_2, x_3, x_4] \quad (1)$$

$$\mathbf{F}\{\boldsymbol{\omega}\} = [\omega_1, \omega_2, \omega_3, \omega_4] \quad (2)$$

$$y_p = \mathbf{F}\{\boldsymbol{\omega}\} \cdot \mathbf{F}\{\mathbf{x}\} + b \quad (3)$$

$$Y_p = y_p \times (N - M) + M \quad (4)$$

上述公式中,以年龄区间为(M, N)的样本训练集构建 SVR 回归模型,其权重向量为 $\mathbf{F}\{\boldsymbol{\omega}\}$,偏置项为 b 。当输入特征向量 $\mathbf{F}\{\mathbf{x}\}$,可获得与预测年龄标准值对应的预测值 y_p ,再对 y_p 进行反归一化处理得到实际年龄预测值 Y_p 。其中 x_1, x_2, x_3, x_4 分别对应 CpG2、CpG3、CpG4、CpG5 位点的甲基化水平。

为确保 SVR 模型在不同参数组合下达到最佳

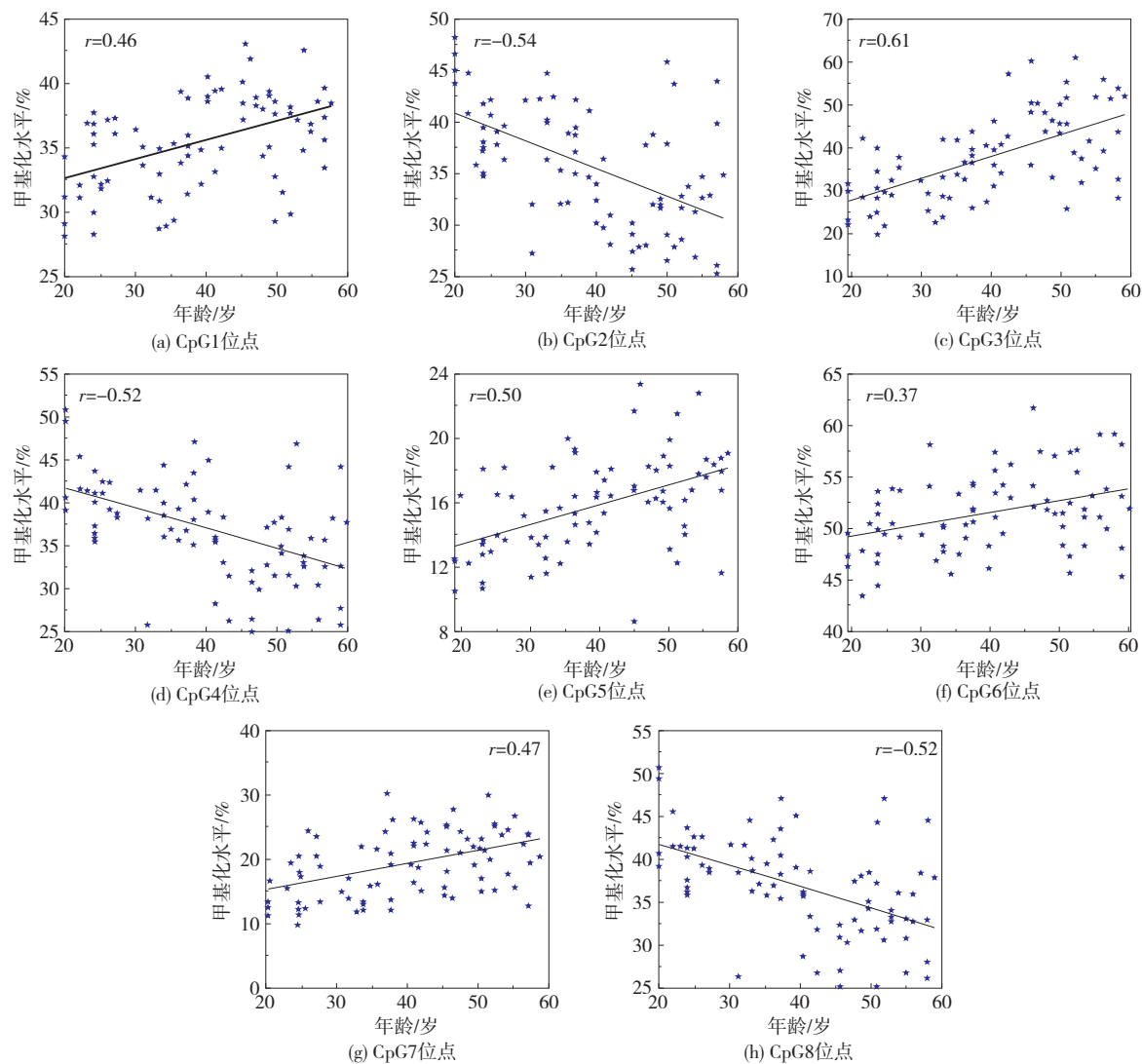


图 1 血液样本 8 个 CpG 位点的 DNA 甲基化水平与样本年龄的回归散点图

Fig. 1 Regression scatter plots between DNA methylation levels and age at 8 CpGs in blood samples

表 3 候选位点对应 r 值

Table 3 Pearson correlation coefficients of candidate CpG sites	
位点编号	r 值
CPG1	0.46
CPG2	-0.54
CPG3	0.61
CPG4	-0.52
CPG5	0.50
CPG6	0.37
CPG7	0.47
CPG8	-0.52

预测性能,本文采用网格搜索(grid search)结合 5 折交叉验证的方法,在设定范围内对 C 和 g 进行

优化,核函数选用径向基函数(radial basis function, RBF),该核函数是目前 SVR 中最常用于捕捉非线性关系的一类。通过逐一测试所有参数组合,计算交叉验证下的平均均方误差(mean squared error, MSE),选出使预测误差最小的参数对。最终确定 $C = 1.41, g = 8.00$ 。该方法能够在避免过拟合的同时提高模型泛化能力,使所建模型在法医学实践中的预测更为稳健可靠。训练模型参数与公式变量值见表 4。

计算结果显示,所建 SVR 回归模型训练集的决定系数(R^2)为 0.908 08,测试集 R^2 为 0.885 29,表明基于 CpG2-CpG3-CpG4-CpG5 位点组合构建的年龄推断模型具有较好的拟合优度。训练集预测年龄与真实年龄的 MAD 为 1.58 岁,测试集的为 2.57 岁。

表4 训练模型参数与公式变量值

Table 4 Training model parameters and variables in formulas

参数/变量	取值	含义说明
t	2	径向基函数(RBF)核
C	1.41	惩罚参数
g	8.00	核函数参数 gamma
s	3	ε -SVR
p	0.01	损失函数中的 epsilon 值
ω_1	-0.95	CpG2的权重
ω_2	0.84	CpG3的权重
ω_3	-0.83	CpG4的权重
ω_4	0.78	CpG5的权重
b	-0.41	回归模型偏置项
M	20	设定样本区间的左边界
N	60	设定样本区间的右边界

训练集和测试集预测年龄与真实年龄的拟合折线图分别见图2和图3。

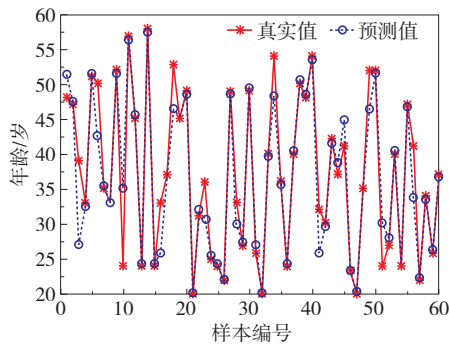


图2 训练集预测年龄与真实年龄的拟合折线图

Fig. 2 Fitting line graph between predicted age and actual age in the training set

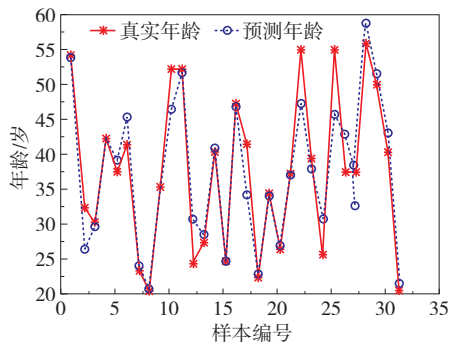


图3 测试集预测年龄与真实年龄的拟合折线图

Fig. 3 Fitting line graph of predicted age and actual age in the test set

2.4 模型在不同年龄组的准确性分析

将91份血液样本中的训练集(样本数为60)和测试集(样本数为31)分别划分成4个年龄组(20~29岁、30~39岁、40~49岁、50~59岁),计算各组的MAD值和模型准确率(以预测偏差在 ± 5 岁范围内为准确)。训练集和测试集各年龄组的准确性指标分别如表5、表6所示。

表5 训练集年龄组别准确性指标

Table 5 Accuracy indicators of the training set by age group

年龄阶段/岁	MAD值	准确率/%	样本数量/个
20~29	1.25	94	17
30~39	2.41	89	18
40~49	2.98	84	13
50~59	3.15	83	12

表6 测试集各年龄组准确性指标

Table 6 Accuracy indicators of the test set by age group

年龄阶段/岁	MAD值	准确率/%	样本数量/个
20~29	2.33	100	9
30~39	2.98	100	8
40~49	3.05	75	7
50~59	3.68	57	7

考虑到样本数量较少对准确率计算的影响,在模型自检的条件下,基本可将误差控制为 ± 5 岁以内。同时也可看到,训练集和测试集的MAD值均随年龄组年龄的增大而增大。

2.5 模型性能分析与改进方向

近年来,在DNA甲基化年龄推断研究中,采用少量强相关CpG位点进行模型构建逐渐成为一种新的策略^[18-19]。本研究通过文献调研筛选出8个与年龄相关的CpG位点,经Pearson相关性分析,发现CpG2、CpG3、CpG4、CpG5这4个位点的 r 值较优,甲基化水平和年龄呈现出较好的线性关系且离散点较少,满足使用SVR建模的要求,适合本研究进行模型构建。

经训练与筛选,本文利用91个Illumina 850K芯片技术检测的血液样本数据,通过SVR建立数学模型,并使用网格搜索与交叉验证寻找最佳参数 C 和 g ,最终构建出的模型测试集MAD为2.57岁,准确率为89%。在此构建方式下,模型准确率与稳定性均较高,未出现过拟合的现象,模型测试集决定系数 R^2 为0.885 29。同时,对测试集的年龄组样本的

评估显示,该模型在不同年龄段均具有较好的重复性和较高的准确率,可将预测误差控制在5岁以内。但是需要注意的是,无论测试集还是训练集的MAD值均随年龄组年龄的增大而增大,说明随着样本年龄增加,模型对未知样本的预测能力有所下降。模型预测误差随样本年龄增长而增大的现象可能与多个因素有关。一方面,CpG位点的甲基化水平在高年龄群体中可能出现非线性变化或趋于饱和,使得SVR模型难以准确捕捉变化趋势;另一方面,高年龄组样本数量相对较少,训练数据在该年龄段的代表性不足,可能影响模型对该人群的拟合效果。此外,疾病、生活方式等外源性因素在高年龄群体中对甲基化水平的影响变异性更大,也可能增加预测误差。未来可考虑引入非线性回归核函数或采用年龄分层建模策略,以提升整体预测精度。

3 结论

本研究基于91份血液DNA甲基化数据,在文献调研与Pearson相关性分析基础上,从8个候选CpG位点中选取CpG2、CpG3、CpG4和CpG5共4个位点,构建了基于SVR的年龄预测模型。结果显示,该模型训练集MAD为1.58岁,测试集MAD为2.57岁,测试集决定系数 R^2 为0.885 29,表明模型具有较好的拟合性能和预测精度。

研究结果表明,仅使用少量与年龄高度相关的CpG位点,即可实现对血液样本年龄的较准确推断,说明低维度位点组合在法医年龄推断中具有可行性和应用潜力。进一步的年龄分组分析显示,该模型在20~39岁青年和中年群体中的预测表现相对更为稳定,而随着年龄增加,预测误差呈上升趋势,高年龄组样本的预测准确性有所下降。

总体而言,本研究建立的血液DNAm年龄预测模型具有较好的稳定性和一定的法医学应用价值,可为犯罪嫌疑人生理特征刻画、身份识别及缩小侦查范围提供技术参考。但本研究仍存在样本量有限、年龄分布范围较窄及高年龄组样本预测效果相对不足等局限。后续研究应进一步扩大样本规模,优化模型参数与年龄分层策略,并加强与法医学实务应用场景的结合,以提升模型的稳定性、适用性和推广价值。

参考文献:

- [1] LU A T, QUACH A, WILSON J G, et al. DNA methylation grim age strongly predicts lifespan and healthspan

- [J]. *Aging*, 2019, 11(2): 303–327.
- [2] CHRISTENSEN B C, HOUSEMAN E A, MARSH C J, et al. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context[J]. *PLoS Genetics*, 2009, 5(8): e1000602.
- [3] FRAGA M F, BALLESTAR E, PAZ M F, et al. Epigenetic differences arise during the lifetime of monozygotic twins[J]. *Proceedings of the National Academy of Sciences of the United States*, 2005, 102(30): 10604–10609.
- [4] HORVATH S. DNA methylation age of human tissues and cell types[J]. *Genome Biology*, 2013, 14(10): 1–20.
- [5] HORVATH S, RAJ K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing[J]. *Nature Reviews Genetics*, 2018, 19(6): 371–384.
- [6] HANNU G, GUINNEY J, ZHAO L, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates[J]. *Molecular Cell*, 2013, 49(2): 359–367.
- [7] SMITH B H, CAMPBELL A, LINKSTED P, et al. Cohort profile: generation Scotland: Scottish family health study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness[J]. *International Journal of Epidemiology*, 2013, 42(3): 689–700.
- [8] ZBIEĆ-PIEKARSKA R, SPÓLNICKA M, KUPIEC T, et al. Development of a forensically useful age prediction method based on DNA methylation analysis[J]. *Forensic Science International: Genetics*, 2015, 17: 173–179.
- [9] SMITH Z D, MEISSNER A. DNA methylation: roles in mammalian development[J]. *Nature Reviews Genetics*, 2013, 14(3): 204–220.
- [10] JONES P A, TAKAI D. The role of DNA methylation in Mammalian epigenetics[J]. *Science*, 2001, 293(5532): 1068–1070.
- [11] LEVINE M E, LU A T, QUACH A, et al. An epigenetic biomarker of aging for lifespan and healthspan [J]. *Aging*, 2018, 10(4): 573.
- [12] LIU C, MARIONI R E, HEDMAN Å K, et al. A DNA methylation biomarker of alcohol consumption [J]. *Molecular Psychiatry*, 2018, 23(2): 422–433.
- [13] HAO T, GUO J L, LIU J D, et al. Predicting human age by detecting DNA methylation status in hair[J]. *Electrophoresis*, 2021, 42(11): 1255–1261.
- [14] BELSKY D W, MOFFITT T E, COHEN A A, et al. Eleven telomere, epigenetic clock, and biomarker-composite quantifications of biological aging: do they measure the same thing?[J]. *American Journal of Epidemiology*, 2018, 187(6): 1220–1230.
- [15] XU C, QU H Z, WANG G Y, et al. A novel strategy for forensic age prediction by DNA methylation and support

- vector regression model [J]. *Scientific Reports*, 2015, 5(1): 17788.
- [16] FENG L, PENG F D, LI S, et al. Systematic feature selection improves accuracy of methylation-based forensic age estimation in Han Chinese males [J]. *Forensic Science International: Genetics*, 2018, 35: 38–45.
- [17] PAN C, YI S H, XIAO C, et al. The evaluation of seven age-related CpGs for forensic purpose in blood from Chinese Han population [J]. *Forensic Science International: Genetics*, 2020, 46: 102251.
- [18] FREIRE - ARADAS A, PHILLIPS C, LAREU M V. Forensic individual age estimation with DNA: from initial approaches to methylation tests [J]. *Forensic Science Review*, 2017, 29(2): 121–144.
- [19] 林培双. 基于国产微滴数字PCR的法医DNA应用研究[D]. 贵阳: 贵州医科大学, 2023.
- LIN P S. Research on the application of droplet digital PCR technology in forensic DNA analysis based on domestic platforms [D]. Guiyang: Guizhou Medical University, 2023. (in Chinese)

A blood DNA methylation age-prediction model based on support vector regression

LIU Rui ZHANG Jing MI HaoYuan FAN Pu

(School of Criminal Investigation, People's Public Security University of China, Beijing 100038, China)

Abstract: Based on blood DNA methylation data obtained from the Illumina Methylation EPIC (850K) Bead Chip, this study aims to establish a highly stable and accurate age-prediction model has been developed using the principles of support vector regression (SVR) with a low-dimensional and high-correlation site selection approach. Four core CpG sites were selected through a systematic literature review and Pearson correlation analysis. The SVR model was developed in Matlab platform using the libsvm toolkit. To determine the optimal parameter combination, grid search combined with cross-validation was employed to optimize the penalty parameter C and the kernel parameter g . During data processing, DNA methylation β -values were normalized for model training and subsequently denormalized to obtain predicted age values. The results demonstrate that the constructed model achieved a coefficient of determination (R^2) of 0.885 29 and a mean absolute deviation (MAD) of 2.57 years on the test set, indicating high overall predictive accuracy. Although the prediction error increased with age—leading to a slight decrease in accuracy for older age groups—the error remained within an acceptable range. This study demonstrates that the SVR model can achieve high-precision age estimation using only four CpG sites, demonstrating robust generalization capability and practical utility. This method provides effective technical support for the physiological characterization of criminal suspects in forensic practice.

Key words: forensic genetics; support vector regression (SVR); age prediction; DNA methylation (DNAm)

(责任编辑:李 蕾)