DOI:10.19603/j.cnki.1000-1190.2026.02.013 **文章编号**:1000-1190(2026)02-0308-13

基于机器学习和氨基酸位置相关系数法的HPV进化关系和亚型分类研究

胡画霖,何黎黎*,刘茂省*

(北京建筑大学理学院,北京 102616)

摘 要:本研究提出了一种基于氨基酸位置相关信息的非序列比对方法——氨基酸相关位置系数法(ACCFV),用于人类乳头瘤病毒(HPV)的进化分析和亚型分类.传统多序列比对方法(MSA)在处理大规模数据时面临计算效率低和内存消耗大的问题,而ACCFV方法通过构建氨基酸之间的位置相关统计量,将氨基酸序列转化为数字特征向量,有效克服了这些限制.研究选取HPV的八种蛋白(E6、E7、E1、E2、E4、E5、L1和L2)的氨基酸序列作为目标数据,利用ACCFV方法提取特征后,通过特征向量间的欧氏距离构建系统进化树,并结合4种机器学习模型进行分类预测.结果显示,当延迟步长$L=1$时,ACCFV方法在进化分析中与传统多序列比对方法Muscle结果高度一致,同时显著提升了计算效率,且随机森林模型的分类准确率达到100%.与BLAST-Protein相比,ACCFV在保持100%分类准确率的同时,处理时间显著缩短,且无需分批操作.本研究不仅验证了ACCFV方法在HPV研究中的可行性和有效性,也为其他病毒的分子流行病学研究提供了新的技术思路.

关键词:HPV;氨基酸序列;机器学习;进化分析;亚型分类

中图分类号:Q811.4 **文献标识码**:A **开放科学(资源服务)标识码(OSID)**:

在全球公共卫生领域,人类乳头瘤病毒(human papillomavirus, HPV)作为一种广泛传播的病原体,长期以来一直是科研界及医学界关注的焦点.HPV是一种具有双层衣壳的球形DNA病毒,广泛存在于自然界中,且以人为唯一的自然宿主[1].这种病毒展现出极强的环境耐受性,特别是在干燥环境中能够长时间保持活性,从而增加其传播的风险[2-3].

HPV病毒家族庞大且复杂,目前已鉴定出超过200种不同的亚型,这些亚型在生物学特性、致病性以及对宿主的影响等方面存在显著差异[4].根据HPV引起的疾病的严重程度,研究人员将其分为高危型和低危型两大类.高危型HPV主要和宫颈癌前病变、宫颈癌等生殖道恶性肿瘤相关[5].高危型HPV,特别是HPV 16型和18型,是宫颈癌发生的主要元凶,约占所有HPV相关宫颈癌和头颈癌的70%[6].除此之外还有HPV 31型、HPV 33型、HPV 35型、HPV 39型、HPV 45型、HPV 51型、HPV 52型、HPV 56型、HPV 58型、HPV 59型、HPV 68型、HPV 73型和HPV 82型等[7-10].低危型HPV一般会使皮肤或黏膜产生良性病变,如寻常疣、扁平疣、尖锐湿疣等.这些亚型包括:HPV 6型、HPV 11型、HPV 32型、HPV 42型、HPV 43型、HPV 44型[11-13].

HPV的致癌机制涉及多个分子过程,其中E6和E7蛋白是病毒编码的两个关键致癌蛋白,在恶性转化中发挥核心作用[14].E6蛋白通过结合宿主细胞内的多种蛋白,抑制细胞凋亡并调节肿瘤抗原表达[15],同时干扰细胞周期调控,促进异常增殖与恶性转化.E7蛋白则主要通过与视网膜母细胞瘤蛋白(retinoblastoma protein,Rb)等关键调控因子相互作用,破坏细胞周期正常进程,诱发肿瘤发生[16].

在病毒分类与进化研究中,传统方法主要依赖于序列比对技术.多序列比对(multiple sequence alignment, MSA)在生物序列的结构与功能分析中扮演关键角色,能够为序列家族的系统发育关系

收稿日期:2025-05-10.
基金项目:国家自然科学基金项目(12571522);北京建筑大学高层次人才引进资助计划项目(GDRC20220802);2024年度北京市数字教育研究课题(青年课题)(BDEC2024QN081);北京市教育委员会2024年度科研计划一般项目(KM202410016001);2024年北京市高等教育学会课题(MS2024130).
*通信联系人. E-mail:liumaoxing@bucea.edu.cn;lilyhe6@163.com.

和功能关联提供重要依据[17]. 诸如Clustal W等工具被广泛应用于蛋白质结构预测、系统发育推断及序列分析，此外还有PREFAB、SABMARK、OXBENCH和IRMBASE等常用基准数据库支持这一过程[18]. MAFFT采用快速傅里叶变换加速同源区域识别，在维持较高精度的同时显著降低了计算时间[19]. T-Coffee则是一种基于树的一致性目标函数的多功能MSA方法，能够整合不同比对策略以及结构、进化或实验信息，从而获得更准确和生物学意义更丰富的比对结果[20-21]. 然而，这类方法在处理大规模数据时普遍面临计算效率低和可扩展性不足的局限.

近年来，随着机器学习技术的发展，非序列比对方法在生物信息学领域中日益广泛应用. 例如，DeepMSA2通过生成高质量多序列比对提升了蛋白质结构预测的准确性[22]；pLM-BLAST在保持与HHsearch相当精度的同时大幅提高了序列搜索速度[23]；MMseqs2则实现了比PSI-BLAST更高的灵敏度，且运行效率提升了数百倍[25]. 这类方法通常将任务划分为编码、特征提取和相似性计算等模块，为大规模序列数据处理提供了有效支撑[24-26]. 特别在面对海量病毒序列数据时，传统的系统发育分析方法往往因计算成本高昂而难以适用，而此时基于机器学习的非序列比对方法能够高效、准确地进行亚型分类，显示出显著优势.

相较于DNA序列比对，氨基酸序列比对在某些方面表现出更为突出的优点，蛋白质可以传递更多信息，在同源序列识别具有更高的准确性[27]. 蛋白质由氨基酸组成，而氨基酸具有多种生物物理和化学性质，如酸碱性、疏水性、亲水性等. 这些性质在蛋白质的功能和结构中起着关键作用. DNA的遗传密码具有简并性，即多个密码子可能编码同一种氨基酸[28]. 这意味着DNA序列中的某些变化可能不会改变其编码的氨基酸序列，从而降低了DNA序列比对在揭示蛋白质功能差异方面的敏感性. 相比之下，氨基酸序列比对能够更准确地反映这些变化对蛋白质功能的影响.

因此，本研究提出了一种基于机器学习的非序列比对方法——氨基酸位置相关系数法（amino acid correlation coefficient feature vector，ACCFV），这种方法通过提取E6、E7、E1、E2、E4、E5、L1和L2共8种蛋白的氨基酸序列特征，形成特征向量，并利用机器学习算法进行训练和测试，从而实现对HPV亚型的系统发育树构建和快速准确分类. 初步研究结果显示，这种方法在HPV进化分析和亚型分类方面表现出色，具有很高的准确性和可靠性.

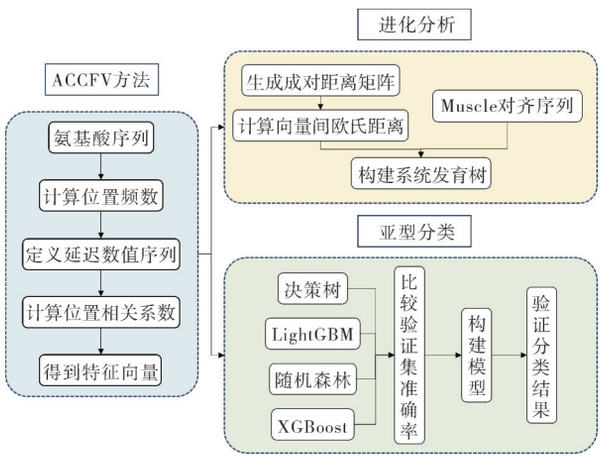

图1　本研究流程图

Fig. 1　Flowchart of the study

1　数据集

1.1　数据预处理

本研究采用的HPV序列数据来源于GitHub公共数据库HPV-Ref-Genomes（https://github.com/vtrevino/HPV-Ref-Genomes）[29]，共获取6 538条HPV全基因组DNA序列. 考虑到氨基酸序列在进化保守性和同源性识别方面的优势，本研究通过以下流程构建蛋白质序列数据集：首先，基于获得的DNA序列的Accession number，从NCBI数据库（https://www.ncbi.nlm.nih.gov/）下载对应的E6、E7、E1、E2、E4、E5、L1和L2蛋白的氨基酸序列. 随后进行数据质量控制，包括序列完整性检查和仅保留包含7条以上序列的HPV亚型，以保证后续统计分析的可靠性.

经过上述筛选，最终获得包含26个HPV亚型、共计33 222条高质量氨基酸序列的数据集（表1）. 该数据筛选策略有效平衡了序列多样性和统计分析效能，为后续基于氨基酸特征的分类研究提供了可靠基础. 其中，E7蛋白的氨基酸序列数量最多（5 276条），反映了其在HPV分型中的标志性作用；而E4蛋白的氨基酸序列相对较少（2 105条）.

1.2　数据集划分

本研究采用分层抽样方法，按照HPV亚型对数据集进行划分，将每个HPV亚型的数据随机分配为75%的训练集和25%的独立测试集，确保训练集和测试集中各亚型样本的比例与其原始分布保持一致. 由于训练集采用五折交叉验证进行模

表1 E6、E7、E1、E2、E4、E5、L1和L2蛋白对应的HPV类型条数

Tab. 1 Number of HPV types associated with each viral protein (E6, E7, E1, E2, E4, E5, L1, and L2)

	亚型	E6	E7	E1	E2	E4	E5	L1	L2
α_3	HPV 61	9	9	9	9	9	0	9	9
	HPV 51	22	22	22	22	22	0	22	22
α_5	HPV 69	7	7	7	7	7	7	7	7
	HPV 82	20	20	20	20	20	20	20	20
α_6	HPV 30	15	15	15	15	14	0	15	15
	HPV 53	24	24	23	24	22	8	24	24
	HPV 56	7	7	0	7	0	0	7	7
	HPV 66	12	12	12	12	12	0	12	12
	HPV 18	119	119	119	119	119	119	44	119
α_7	HPV 39	19	19	19	19	19	19	19	19
	HPV 45	12	12	12	12	0	0	12	12
	HPV 59	7	7	8	8	8	8	7	8
	HPV 68	19	19	20	20	0	20	20	20
	HPV 70	9	9	8	8	9	9	8	8
α_9	HPV 16	2 370	3 498	2 883	3 453	412	3 288	410	3 056
	HPV 31	24	24	24	24	24	24	24	24
	HPV 33	23	23	23	23	23	23	22	23
	HPV 35	845	897	897	867	868	874	896	895
	HPV 52	82	82	82	82	80	80	82	82
	HPV 58	135	135	135	135	135	135	134	135
	HPV 67	7	7	7	7	7	7	7	7
α_{10}	HPV 6	185	185	187	187	187	187	187	187
	HPV 11	87	87	87	87	87	87	86	87
α_{11}	HPV 34	15	15	15	15	0	14	15	15
	HPV 73	12	12	12	12	11	11	12	12
α_{13}	HPV 54	9	10	10	10	10	0	10	10
	总计	4 095	5 276	4 656	5 204	2 105	4 940	2 111	4 835

型训练和参数优化,因此未设置独立的验证集.这一划分策略既保证了模型训练的充分性,又确保了测试集评估结果的客观性.为保障实验的可重复性,研究设置了固定的随机种子(random_state=42),同时对样本量较少的亚型采用过采样技术以提高其在训练集中的代表性.

1.3 去重策略

为避免数据泄露并确保模型的泛化能力,本研究实施了严格的数据去重流程.首先,对训练集和测试集中的重复序列进行全局去重,即移除所有完全相同的氨基酸序列,仅保留每条唯一序列的一条记录.其次,通过序列比对确保独立测试集中的任何序列均未出现在训练集中,从而消除因序列重叠导致的评估偏差.经过去重处理后,各蛋白数据集的序列数量有所减少,但分类性能仍保持高度稳定.具体去重后的数据集规模及分类结果(包括准确率、$F1$-score和Recall值)详见文后附录A,结果显示随机森林模型在去重后的数据集上依然能够实现99%以上的分类准确率,进一步验证了ACCFV方法的鲁棒性和可靠性.

2 方法

2.1 ACCFV法

ACCFV法通过提取20种氨基酸在序列中的位置相关性信息,进一步将序列转化成数字特征向量,以供机器学习模型使用[30].给定氨基酸序列$P=p_1p_2\cdots p_N$, $p_i\in\{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}$,氨基酸$\phi=\{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y\}$,首先通过如下的示性函数把氨基酸序列转化为20条0~1序列,其中对应氨基酸ϕ的序列为$X_\phi=\big(x_\phi(1),x_\phi(2),\cdots,x_\phi(N)\big)$,示性函数为:

$$x_\phi=\begin{cases}1, p_i=\phi,\\ 0, \text{其他}.\end{cases}\tag{1}$$

计算位置频数:

$$f_\phi=\frac{1}{N}\big(x_\phi(1)+x_\phi(2)+\cdots+x_\phi(N)\big),\tag{2}$$

接下来定义一个L步长的延迟数值序列:

$$X_{\phi+L}=\big(x_\phi(L+1),x_\phi(L+2),\cdots,x_\phi(L+N)\big).\tag{3}$$

记氨基酸ϕ,ω之间的位置相关关系为:

$$\rho_{\phi\omega}(L)=\frac{1}{N}\sum_{i=1}^{N}\big(x_\phi(i)-f_\phi\big)\big(x_{\omega+L}(i)-f_\omega\big).\tag{4}$$

特别地:

$$\rho_{\phi\phi}(0)=\frac{1}{N}\sum_{i=1}^{N}\big(x_\phi(i)-f_\phi\big)\big(x_{\phi+L}(i)-f_\phi\big).\tag{5}$$

因此,得到氨基酸ϕ,ω之间的位置相关系数为:

$$\tau_{\phi\omega}(L)=\frac{\rho_{\phi\omega}(L)}{\sqrt{\rho_{\phi\phi}(0)\rho_{\omega\omega}(0)}},\tag{6}$$

特别地:

$$\tau_{\phi\phi}(L)=\frac{\rho_{\phi\phi}(L)}{\rho_{\phi\phi}(0)}.\tag{7}$$

最终,把所有的位置相关系数放在一起形成一个$400\times L$维的向量$V=(\tau_{AA}(1),\tau_{AA}(2),\cdots,\tau_{AA}(L),\tau_{AC}(1),\tau_{AC}(2),\cdots,\tau_{AC}(L),\cdots,\tau_{AY}(1),\tau_{AY}(2),\cdots,\tau_{AY}(L),\tau_{CA}(1),\tau_{CA}(2),\cdots,\tau_{CA}(L),\tau_{CC}(1),\tau_{CC}(2),$

$\cdots,\tau_{CC}(L),\cdots,\tau_{CY}(1),\tau_{CY}(2),\cdots,\tau_{CY}(1),\cdots,\cdots,$ $\tau_{YY}(1),\tau_{YY}(2),\tau_{YY}(L))$.

例：给定序列 MESANASTPA，下面给出 ACCFV($L=1$)的详细计算流程(图2)：

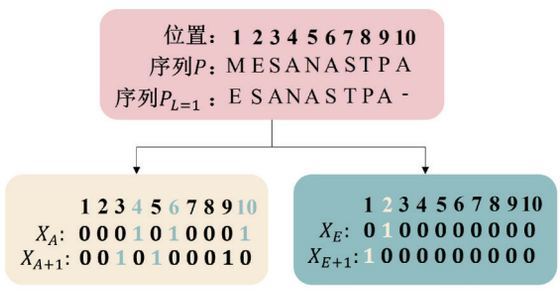

图2　示例序列的位置系数图解

Fig. 2　Positional correlation coefficient diagram of the example sequence

$$X_A=[0,0,0,1,0,1,0,0,0,1],$$
$$X_{A+1}=[0,0,1,0,1,0,0,0,1,0],$$
$$X_E=[0,1,0,0,0,0,0,0,0,0],$$
$$X_{E+1}=[1,0,0,0,0,0,0,0,0,0].$$
$$\rho_{AA}(0)=\frac{1}{10}\left[\left(1-\frac{3}{10}\right)\times 3\right]=\frac{21}{100},$$
$$\rho_{EE}(0)=\frac{1}{10}\left(1-\frac{1}{10}\right)=\frac{9}{100},$$
$$\rho_{AA}(1)=\frac{1}{10}\left[\left(1-\frac{3}{10}\right)\left(0-\frac{3}{10}\right)\times 3+\left(0-\frac{3}{10}\right)\left(0-\frac{3}{10}\right)\times 4+\left(0-\frac{3}{10}\right)\left(1-\frac{3}{10}\right)\times 3\right]=-\frac{9}{100},$$
$$\rho_{EE}(1)=\frac{1}{10}\left[\left(0-\frac{1}{10}\right)\left(1-\frac{1}{10}\right)+\left(0-\frac{1}{10}\right)\left(0-\frac{1}{10}\right)\times 8+\left(1-\frac{1}{10}\right)\left(0-\frac{1}{10}\right)\right]=-\frac{1}{100},$$
$$\rho_{AE}(1)=\frac{1}{10}\left[\left(0-\frac{3}{10}\right)\left(1-\frac{1}{10}\right)+\left(0-\frac{3}{10}\right)\left(0-\frac{1}{10}\right)\times 6+\left(1-\frac{3}{10}\right)\left(0-\frac{1}{10}\right)\times 3\right]=-\frac{3}{100},$$
$$\tau_{AA}(1)=\frac{\rho_{AA}(1)}{\rho_{AA}(0)}=\frac{3}{7},$$
$$\tau_{EE}(1)=\frac{\rho_{EE}(1)}{\rho_{EE}(0)}=-\frac{1}{9},$$
$$\tau_{AE}(1)=\frac{\rho_{AE}(1)}{\sqrt{\rho_{AA}(0)\rho_{EE}(0)}}=\frac{1}{\sqrt{21}}.$$

接着，依次计算出 $\tau_{AM}(1)$、$\tau_{AN}(1)$、$\tau_{AP}(1)$、$\tau_{AS}(1)$、$\tau_{AT}(1)$、$\tau_{EA}(1)$、$\tau_{EM}(1)$、…、$\tau_{ET}(1)$、$\tau_{MA}(1)$、$\tau_{ME}(1)$、…、$\tau_{MT}(1)$、…、$\tau_{TA}(1)$、$\tau_{TE}(1)$、…、$\tau_{TT}(1)$，则序列“MESANASTPA”被转换成 400 维的数字特征向量.

2.2　模型构建

在ACCFV方法模型构建过程中，首先对延迟步长 L 的取值进行优化选择. 通过将 L 的取值从1滑动到5，并在E7测试集上评估预测准确率(图3)，发现当 $L=1$ 时准确率达到100%，$L=2$ 时略微下降至99.7%，而 L 取3～5时准确率又回升至100%. 基于计算复杂度最小化的原则，最终选定 $L=1$ 作为最优延迟步长.

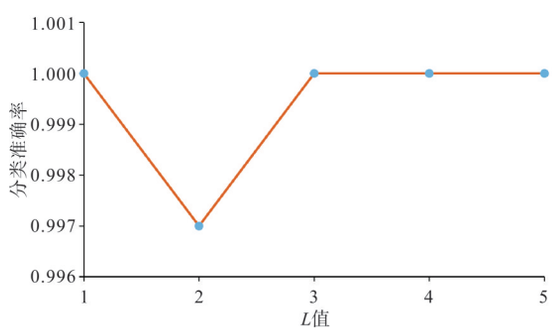

图3　HPV E7测试集上不同 L 参数的预测准确率

Fig. 3　Prediction accuracy of the HPV E7 test set with varying L parameters (L=1-5)

在确定参数 L 后，利用ACCFV方法将氨基酸序列转化为数值向量，并采用4种机器学习方法(决策树、LightGBM、随机森林和XGBoost)进行HPV亚型分类建模[32]. 模型训练过程中采用五折交叉验证进行评估，并以独立测试集的预测准确率作为最终评价指标来选择最优模型. 为确保实验的完全可重复性，在数据分层抽样、五折交叉验证的折划分以及所有包含随机过程的机器学习模型中，均设置了固定的随机种子. 以E7蛋白数据集为例，随机森林模型在独立测试集上对所有类型的预测准确率均达到1(表2)，表现出完美的分类性能. 其他数据集的预测结果详见文后附录B，综合分析后，最终选择参数 $L=1$ 的随机森林模型作为ACCFV方法的分类模型.

表2 ACCFV(L=1)方法的四种机器学习模型在E7蛋白测试集上的预测准确率

Tab. 2 Performance comparison of four machine learning with ACCFV (L=1) on HPV E7 protein test set

	6	11	16	18	30	31	33	34	35	39	45	51	52
决策树	1.00	1.00	1.00	1.00	0.96	0.90	0.92	0.87	1.00	1.00	1.00	0.96	1.00
LightGBM	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.87	1.00	1.00	1.00	1.00	0.99
随机森林	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
XGBoost	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
	53	54	56	58	59	61	66	67	68	69	70	73	82
决策树	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.67	0.93	1.00	1.00	0.88	1.00
LightGBM	1.00	1.00	0.78	1.00	1.00	1.00	1.00	0.83	1.00	1.00	1.00	1.00	1.00
随机森林	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
XGBoost	1.00	0.88	1.00	1.00	1.00	1.00	1.00	1.00	0.93	1.00	1.00	1.00	1.00

2.3 进化分析

给定数据量为N的数据集,通过ACCFV方法和2.2模型构建选出的参数L可以把每一条氨基酸序列P转化为$400\times L$的向量$\boldsymbol{V}$,为此,整个数据集可以转化为一个$n\times(400\times L)$的矩阵$M_{n\times(400\times L)}$. 接着采用欧氏距离的方法,计算$\boldsymbol{M}$中两两序列的欧氏距离,得到距离矩阵$\boldsymbol{D}_{n\times n}$,其元素定义为:

$$\boldsymbol{D}_{ij}=\sqrt{\sum_{k=1}^{400}\left(V_{ik}-V_{jk}\right)^2}, \tag{8}$$

接着将距离矩阵$\boldsymbol{D}$导入MEGA11[31]软件,采用邻接法(neighbor joining)[33]构建系统发育树,最终得到HPV氨基酸序列间的进化关系.

用同一个数据集导入MEGA11软件,采用Muscle算法[34]对HPV E6和E7蛋白的氨基酸序列进行多序列比对,参数设置:gap open为−2.9,gap extend为0,hydrophobicity multiplier为1.2,clustering method为UPGMB,min diag lenth为24. 接着使用邻接法构建系统进化树,参数设置为:Poisson模型,成对缺失数据处理,1 000次bootstrap重复,位点进化速率服从Gamma分布(α=4),得到Muscle算法下HPV的进化关系.

2.4 方法比较

为评估ACCFV方法的性能,本研究从序列比对与亚型分类两个维度将其与现有典型方法进行比较. 在序列比对方面,选用Muscle、Clustal W[35]、T-Coffee及MAFFT作为对比方法. 这些方法基于渐进、迭代或一致性策略,能够通过构建系统发育树清晰呈现病毒进化关系;其中MAFFT在处理大规模数据时仍能保持良好性能,而其余方法更适用于小规模数据的进化分析. 在亚型分类方面,则采用NCBI BLAST-Protein[36]和MMseqs2作为基准,二者基于快速序列比对或聚类策略,适用于大规模数据的高效分类. 通过将ACCFV与涵盖不同计算策略的多种代表性方法进行对比,可从运行效率和分类效果两方面全面评估其性能,从而建立一个多样化且可靠的基准框架. 各方法特点、用途及适用数据规模见表3.

表3 七种方法的特点、用途和数据规模

Tab. 3 Characteristics, applications, and data scale of the seven methods

方法	特点	在本研究中的用途	适用数据规模
ACCFV	氨基酸位置特征提取	进化分析、亚型分类	大规模
Muscle	迭代优化算法,速度快,适用于中等规模数据	进化分析、评估ACCFV运算效率	中等规模
Clustal W	渐进式全局比对算法,是传统经典方法	评估ACCFV运算效率	小规模
T-Coffee	一致性算法,准确性高,但计算资源消耗大	评估ACCFV运算效率	小规模
MAFFT	渐进迭代算法,在速度与精度间有良好平衡	评估ACCFV运算效率	大规模
NCBI BLAST-Protein	基于启发式算法的局部序列比对与数据库搜索工具	验证ACCFV的分类结果准确性	大规模
MMseqs2	极快的大规模序列聚类工具	评估ACCFV运算效率	大规模

2.4.1　序列对齐　在HPV致癌机制中，E6蛋白通过泛素化降解p53蛋白，E7蛋白则与视网膜母细胞瘤蛋白(pRb)结合并使其失活．本研究选取了HPV E6蛋白(4 095条)和E7蛋白(5 276条)的氨基酸序列作为测试数据集，在相同硬件平台(Intel® Core™ i7-13700H CPU@2.40 GHz，16 GB RAM)上比较了Muscle、ClustalW、T-Coffee和ACCFV(基于Python3.11实现)4种方法的运行时间．需要注意的是，MAFFT方法运行于EBI在线服务器(https://www.ebi.ac.uk/Tools/msa/mafft/)，其计算依托于远程高性能计算资源，因此其运行时间是在此特定环境下获得，在此一并列出以供参考．

由于Muscle、ClustalW和T-Coffee难以一次性处理上千条序列，本文采用分层抽样分别从E6和E7数据集中抽取了889条和827条序列构成子集用于对比．所有方法均使用统一数据集：Muscle和ClustalW通过MEGA11软件执行，其中Muscle参数与第2.2节进化分析设置一致，ClustalW参数设为：空位开放罚分10.0，空位扩展罚分0.2，蛋白质权重矩阵为Gonnet，启用残基特异性罚分和亲水罚分，空位分离距离为4，末端空位分离关闭．T-Coffee(https://tcoffee.crg.eu/apps/tcoffee/do:regular)使用在线工具(Version_11.00)以regular模式运行，最大长度10 000，多核数为4，其余参数保持默认．MAFFT通过EBI在线服务运行，参数设置为：矩阵BLOSUM62，空位开放罚分1.53，空位扩展罚分0.123，输出顺序按输入排列，重建树次数为2，输出引导树，最大迭代次数为2，FFT不启用．

2.4.2　亚型分类　为验证ACCFV方法在HPV E6和E7蛋白测试集上的亚型分类效果，本研究选用NCBI BLAST-Protein和MMseqs2作为基准方法进行对比．所有实验均在相同硬件平台(Intel®Core™ i7-13700H CPU@2.40 GHz，16 GB RAM)上运行，ACCFV由Python 3.11实现．

BLAST-Protein适用于高精度单序列数据库搜索，但处理速度较慢．由于在线接口对单次查询的序列总长度有限制，将E6蛋白测试集(1 024条)分为11组，E7蛋白测试集(1 319条)分为14组，每组不超过100条序列提交至BLAST NR数据库，参数为默认设置．对返回的每条序列结果，选取一致性百分比最高且E值最低的匹配[37]作为其预测亚型，以此作为高可靠性标准与ACCFV分类结果进行准确性比较．

MMseqs2适用于快速大规模序列聚类，可能遗漏低相似度功能序列．使用静态编译版，在CMD中运行相同测试集，参数设置为：min-seq-id为0.9，覆盖度阈值(-c)o 0.8，线程数(--threads)为8．由于MMseqs2输出为聚类文件，不便于直接计算分类准确率，因此仅将其进行聚类的时间与ACCFV的运行时间进行对比．

3　结果

3.1　E6和E7蛋白进化分析

按照2.4.1节所述参数，分别对E6和E7蛋白的完整数据集及其抽样子集进行序列处理：使用Muscle、ClustalW、T-Coffee和MAFFT进行多序列比对，同时采用ACCFV方法将氨基酸序列转换为特征向量，5种方法的处理时间如表4所示．

表4　5种方法的处理时间结果表

Tab.4　Computational time of ACCFV, Muscle, ClustalW, T-Coffee and MAFFT methods

数据集	序列数/个	ACCFV/s	Muscle/s	ClustalW/s	T-Coffee/s	MAFFT/s
E6蛋白	4 095	78.32				59.88
	889	15.70	28.82	468.08	1 102.15	13.01
E7蛋白	5 276	69.70				98.67
	827	12.53	29.40	293.07s	1 014.28	16.24

注：MAFFT方法通过EBI在线服务器(https://www.ebi.ac.uk/Tools/msa/mafft/)运行，其背后为高性能计算集群；作为对比，ACCFV及其他对比方法(Muscle，ClustalW，T-Coffee)均运行于本地实验平台，硬件配置为Intel® Core™ i7-13700H CPU@2.40 GHz处理器及16 GB RAM．请注意，运行时间的直接对比可能受到计算资源的显著影响．

Muscle、ClustalW和T-Coffee方法在处理数千条序列时均因超出运算负荷而无法完成比对；MAFFT凭借其内嵌的服务器支持能够有效应对大规模数据，其在E6完整数据集上耗时59.88 s，比ACCFV快18.44 s，但在E7数据集上用时98.67 s，反而比ACCFV多出28.97 s(表4)．在子集上，ACCFV表现出与MAFFT相当的处理效率，其速度约为Muscle的2倍、ClustalW的20倍以上，并达到T-Coffee的100倍以上．值得注意的是，当前ACCFV是在本地终端设备上运行，若能部署至高性能服务器环境，其运算速度还有大幅提升的空间，进一步凸显其高效与可扩展性优势．

基于E6和E7蛋白在HPV致癌机制中的关键作用，本研究选取这两类蛋白作为代表性分子标记

进行进化分析,分别使用ACCFV和Muscle两种方法对同一批序列构建系统发育树(图4~图7).

3.1.1 E6蛋白 针对E6蛋白数据集开展比较进化分析,从每个HPV亚型中随机选取两条代表性序列进行研究.首先,基于ACCFV方法将氨基酸序列转换成特征向量,通过计算向量间的欧氏距离构建成对距离矩阵,采用邻接法(neighbor-joining)构建系统发育树(图4).为验证分析结果的稳健性,本文同时使用传统序列比对方法,通过MEGA11软件中的Muscle算法对相同数据集进行多序列比对,采用邻接法构建系统发育树(图5).

ACCFV方法构建的发育树(图4)显示出清晰的进化关系,其中α_9组群(16、31、33、35、52、58亚型等)形成高度支持的聚类,反映出这些高危型HPV在E6蛋白序列上的保守性.α_7组群(18、39、45亚型等)与α_9组群的进化距离较近,与已知的致癌亚型分类一致,而低危型HPV(如α_{10}组的6、11亚型)则独立成支,与高危型明显分离.相比之下,Muscle方法构建的发育树(图5)虽然在整体组群划分上与ACCFV结果相似,但在分支细节上存在差异,例如部分亚型(如68、73)的定位略有偏移.此外,ACCFV方法在α_5组群(51、82亚型)的分辨率更高,能够清晰区分近缘亚型的进化差异,而Muscle方法在这些细节上的表现相对模糊.两种方法在主要进化关系的呈现上具有一致性,但ACCFV方法在计算效率和细节解析上更具优势,为大规模HPV进化分析提供了更可靠的工具.

3.1.2 E7蛋白 在针对E7蛋白数据集的进化分析中,本研究同样采取从每个HPV亚型中随机选取2条代表性序列的策略,并分别应用ACCFV方法和Muscle算法来构建系统发育树(图6和图7).通过对比这2种方法所构建的发育树,揭示了它们在宏观的进化关系展示上高度的一致性——均清晰地揭示了不同HPV亚型之间的亲缘关系.

值得注意的是,ACCFV方法通过特征向量直接计算序列间距离,避免了多序列比对引入的空位罚分偏差.对长度变异较大的E7蛋白(如HPV 16型序列长度差异达15个氨基酸)仍能保持稳定的距离度量;在保留关键功能域进化信号的同时,减少了非保守区段对整体树形结构的干扰.这些结果验证了ACCFV方法在HPV进化分析中的可靠性,尤其适用于大规模数据集的高效处理.其优势在于将序列信息转化为可量化的特征向量,既克服了传统比对方法的内存限制,又提供了更精细的进化尺度分析能力.

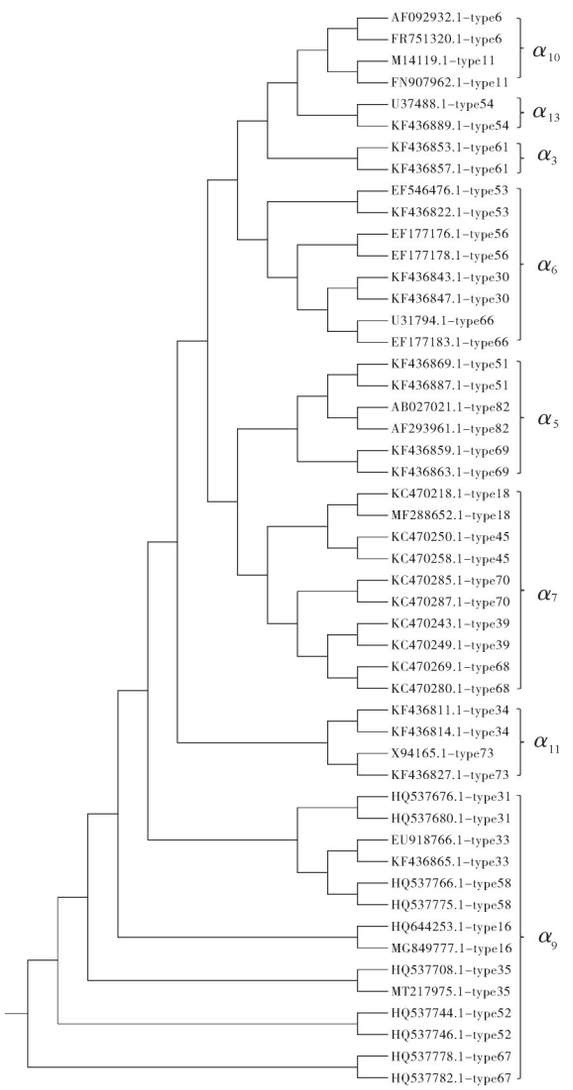

图4 基于ACCFV方法在E6蛋白上的系统发育树

Fig. 4 Phylogenetic tree of E6 protein constructed by ACCFV

3.2 分类结果

在处理大规模序列数据时,传统基于比对的系统发育分析方法常面临效率限制.ACCFV方法从非比对角度出发,通过将氨基酸序列转化为数值特征,有效避免了多序列比对的计算瓶颈,并利用机器学习方法挖掘序列中的进化特征,构建出不依赖于系统发育树的分类框架.这一策略在保持进化信息解析能力的同时,显著提升了处理效率,尤其适用于大规模数据的快速分类.

为评估ACCFV的分类效果,本文将其与BLAST-Protein及MMseqs2进行对比.如表5所示,ACCFV在E6和E7测试集上均达到与BLAST-Protein相同的100%准确率,但耗时显著更短.BLAST-Protein不仅处理速度慢,还需分批

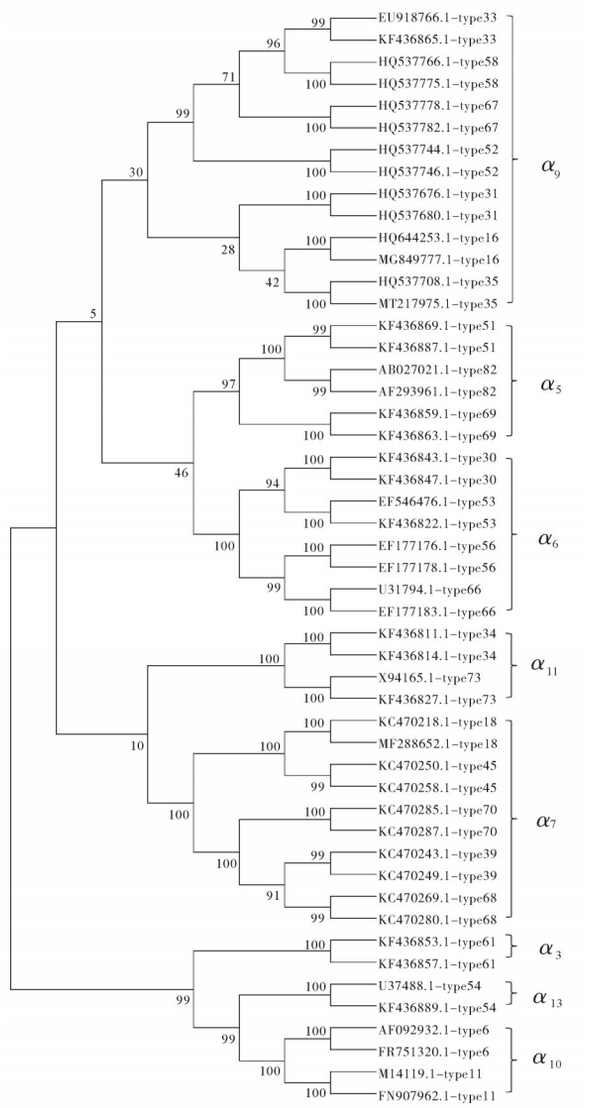

图5　基于Muscle算法在E6蛋白上的系统发育树

Fig. 5　Phylogenetic tree of E6 protein constructed by Muscle

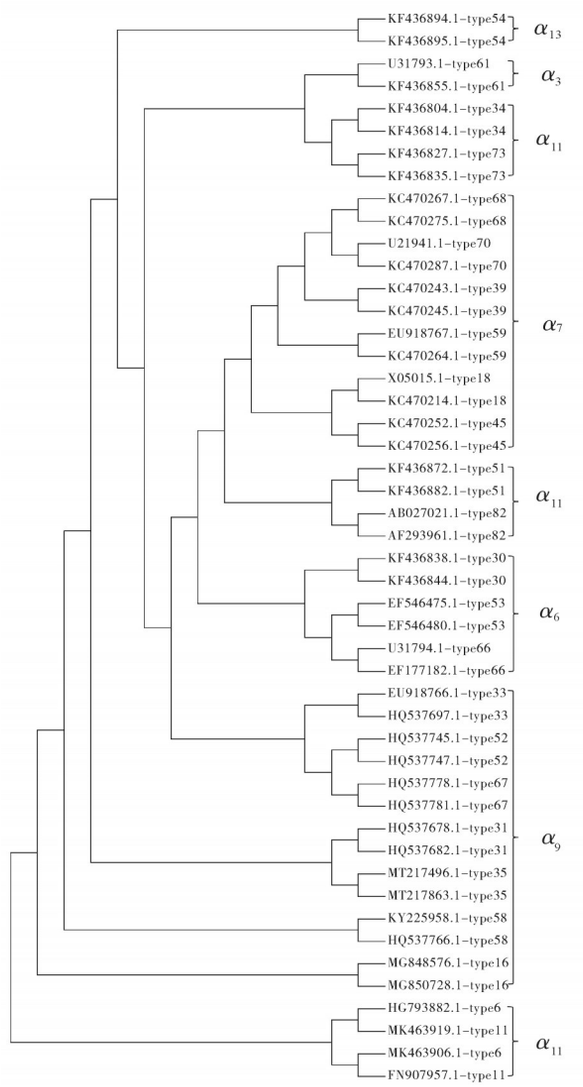

图6　基于ACCFV方法在E7蛋白上的系统发育树

Fig. 6　Phylogenetic tree of E7 protein constructed by ACCFV

操作，易受网络及服务器等因素干扰，不适于高通量场景. MMseqs2虽在速度上与ACCFV接近，但其输出为聚类结果，需借助Cytoscape等工具进行后续分析和亚型判别，操作复杂且依赖使用者的专业知识. ACCFV能够直接输出亚型分类结果，在保证准确性的同时提供更完整的端到端解决方案，兼具高效性与实用性.

在$L=1$的情况下，使用随机森林模型的ACCFV方法在E7蛋白的独立测试集上每一类的预测准确率、F1-score和Recall值如表6所示，从预测准确率来看，ACCFV方法能100%准确预测HPV亚型，F1-score和Recall值也都是1，证明构建的模型可以根据特征向量正确识别所有亚型. 并且数据集中HPV的亚型类型有26种，各个类别的数据量极度不均衡，E7测试集1 319条序列中最少的亚型序列数只有2条，最多的亚型为HPV 16型，

表5　3种方法在E6、E7蛋白测试集上预测结果用时对比

Tab. 5　Comparison of processing time among three methods on E6 and E7 protein test sets

数据集	ACCFV		BLAST-Protein			MMseqs2	
	准确率/%	总时长/s	准确率/%	总时长/s	平均时长(标准差)/s	准确率/%	总时长/s
E6	100	15.70	100	1 101.07	100.1(28.26)		17.32+(亚型判别时间)
E7	100	12.54	100	973.38	69.53(14.82)		12.20+(亚型判别时间)

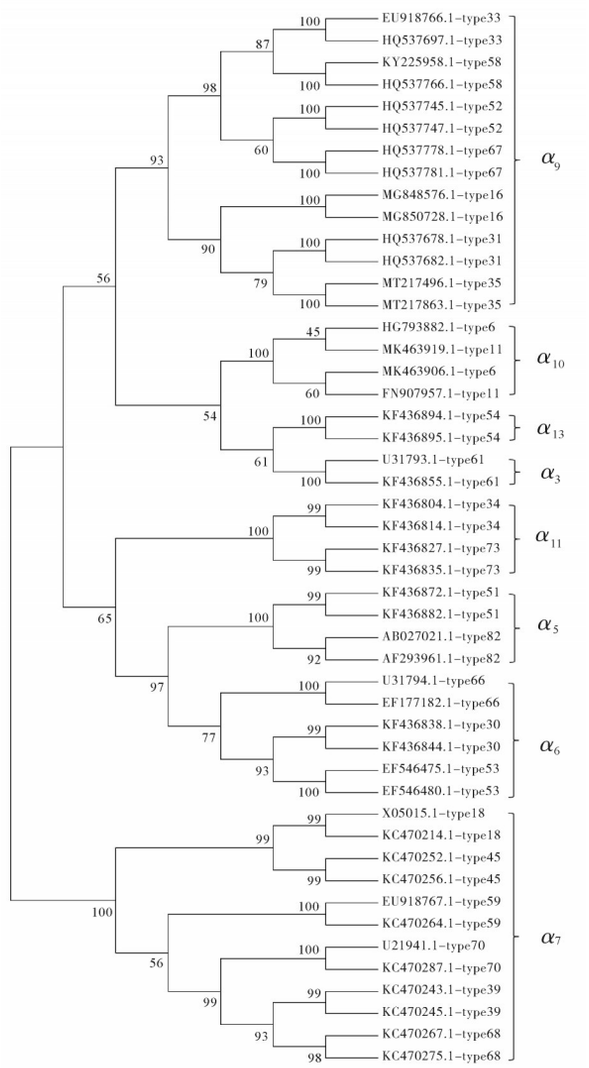

图7 基于Muscle算法在E7蛋白上的系统发育树

Fig. 7 Phylogenetic tree of E7 protein constructed by Muscle

有874条,但基于ACCFV(L=1)的方法仍然能实现准确的亚型分类.

在L=1的情况下,使用随机森林模型的ACCFV方法在E4蛋白的测试集上预测准确率F1-score和Recall值如表7所示,E4测试集527条序列中最少的亚型序列数只有2条,最多的亚型为HPV 35型,有217条,在数据量不均衡的条件下,总体准确率达99%以上,说明ACCFV方法对于氨基酸序列的识别是准确可靠的.其他6种蛋白E6、E1、E2、E5、L1和L2在测试集上的预测结果如附录B所示.

4 结论

本研究提出的氨基酸位置相关系数法创新性地将氨基酸序列转换为蕴含位置信息的数值特征向量,能够有效识别对蛋白质功能或结构具有关键作用的位点,具备明确的生物学意义.以E7蛋白的CR2功能域(pRb结合域)为例[38],分析结果显示,高危型HPV 16与低危型HPV 6呈现显著差异:HPV 16中"天门冬氨酸-亮氨酸"(DL)组合表现出极强的位点协同性(τ_{DL}=0.859),这与该病毒优化pRb结合界面、增强致癌能力的生物学特性高度吻合;而HPV 6中"甘氨酸-亮氨酸"(GL)组合则保持更强的协同模式(τ_{GL}=0.862),与其低致病性特征相一致.这些发现表明,ACCFV方法生成

表6 基于ACCFV(L=1)方法的随机森林模型在E7蛋白测试集上的准确率、F1-score和Recall值

Tab. 6 Performance metrics (accuracy/F1-score/Recall) of RF classifier with ACCFV (L=1) for E7 protein test set

	亚型	准确率/%	F1-score	召回率
α_3	HPV61	100	1.00	1.00
	HPV51	100	1.00	1.00
α_5	HPV69	100	1.00	1.00
	HPV82	100	1.00	1.00
α_6	HPV30	100	1.00	1.00
	HPV53	100	1.00	1.00
	HPV56	100	1.00	1.00
	HPV66	100	1.00	1.00
α_7	HPV18	100	1.00	1.00
	HPV39	100	1.00	1.00
	HPV45	100	1.00	1.00
	HPV59	100	1.00	1.00
	HPV68	100	1.00	1.00
	HPV70	100	1.00	1.00
α_9	HPV16	100	1.00	1.00
	HPV31	100	1.00	1.00
	HPV33	100	1.00	1.00
	HPV35	100	1.00	1.00
	HPV52	100	1.00	1.00
	HPV58	100	1.00	1.00
	HPV67	100	1.00	1.00
α_{10}	HPV6	100	1.00	1.00
	HPV11	100	1.00	1.00
α_{11}	HPV34	100	1.00	1.00
	HPV73	100	1.00	1.00
α_{13}	HPV54	100	1.00	1.00

表7　基于ACCFV(L=1)方法的随机森林模型在E4测试集上的准确率、F1-score和Recall值

Tab. 7　Performance metrics (accuracy/F1-score/Recall) of RF classifier with ACCFV (L=1) for E4 protein test set

	亚型	准确率/%	F1-score	召回率
α_3	HPV61	100	1.00	1.00
	HPV51	100	1.00	1.00
α_5	HPV69	100	1.00	1.00
	HPV82	100	1.00	1.00
α_6	HPV30	100	1.00	1.00
	HPV53	100	1.00	1.00
	HPV56	100	1.00	1.00
	HPV66	100	1.00	1.00
α_7	HPV18	100	1.00	1.00
	HPV39	100	1.00	1.00
	HPV45	100	1.00	1.00
	HPV59	100	1.00	1.00
	HPV68	100	1.00	1.00
	HPV70	100	1.00	1.00
α_9	HPV16	100	1.00	1.00
	HPV31	100	1.00	1.00
	HPV33	100	1.00	1.00
	HPV35	100	1.00	1.00
	HPV52	100	1.00	1.00
	HPV58	100	1.00	1.00
	HPV67	100	1.00	1.00
α_{10}	HPV6	100	1.00	1.00
	HPV11	100	1.00	1.00
α_{11}	HPV34	100	1.00	1.00
	HPV73	100	1.00	1.00
α_{13}	HPV54	100	1.00	1.00

的特征向量并非抽象的数学表征，而是直接反映了蛋白质关键功能域内氨基酸间的进化约束与协同关系，为从序列层面解释HPV型别间致病差异提供可靠的计算生物学依据.

在病毒进化分析与亚型分类研究中展现出全面而卓越的性能. 在进化分析方面，基于ACCFV特征向量构建的系统发育树与经典多序列比对方法(如Muscle)结果高度一致，不仅能够清晰区分高危型HPV(包括HPV 16、31、33、35、52、58亚型等)与低危型HPV(如HPV 6、11亚型)，还能准确反映不同亚型间的系统发育关系，与已知生物学分类完全吻合. 值得注意的是，ACCFV有效规避了多序列比对中因空位罚分设置引起的主观偏差，即使面对长度变异显著的序列(如HPV 16型E7蛋白中存在15个氨基酸的长度差异)，仍能保持进化距离度量的稳定性，从而显著提升了系统发育推断的可靠性.

在分类任务中，当延迟步长参数L=1时，基于ACCFV的随机森林分类器在HPV测试集上准确率超过99%，成功识别全部26种HPV亚型，且不受训练数据中类别不平衡(如某些亚型仅有7条序列，而HPV 16型多达3 498条)的影响. 其他主流机器学习模型(包括决策树、LightGBM和XGBoost)在相同特征上的分类准确率均超过99%，进一步证明了该方法的稳健性与泛化能力. 特别值得关注的是，ACCFV在禽流感病毒数据中也展现出优异的跨物种适用性(详见附录C)：在包含1 163条氨基酸序列的数据集上，随机森林模型取得了100%的准确率，其他模型的准确率均超过97%，同时保持极高的计算效率(运行时间45.66 s，内存占用570.46 MB). 在面对未知亚型时，该方法亦展现出良好的泛化性. 在额外测试的153种新亚型中，仅有部分样本被以极低置信度预测为已知类别(59例为16型，55例为35型，6例为18型)，其最大判定概率仅为0.26，平均概率为0.17，远低于模型对已知亚型所采用的置信阈值0.86. 结果表明，模型对未知样本表现出高度谨慎的判别特性，有效避免了高置信度的错误分类.

ACCFV方法在计算效率方面也展现出显著优势，其处理速度达到传统Muscle方法的两倍以上，并能够高效稳定地处理大规模序列数据. 例如，在包含5 276条E7蛋白序列的数据集上运行期间，最大内存占用始终控制在1 GB以内；即便在更大规模的数据集(33 222条序列)上运行，该方法也仅耗时162.84 s，且内存占用未超过2 GB. 值得注意的是，当前实验受限于计算资源，所有测试均在CPU环境中完成. 若未来部署于GPU加速环境，ACCFV方法的运算效率有望得到进一步提升.

由于20种标准氨基酸的字段重复率较低，且实验表明当延迟步长参数L=1时，模型已能够充分捕获序列中的判别性特征，在保证最高分类精度的同时兼顾计算效率，因此未对参数L进行更大范围的调整. 上述结果表明，ACCFV方法不仅能够

高效处理大规模序列数据,克服传统多序列比对方法在计算效率和内存消耗上的固有局限,还可同时实现高精度亚型分类与稳健的进化分析,为HPV乃至其他病毒的分子流行病学研究提供了一种新的技术路径.

本研究提出的ACCFV方法在HPV亚型分类和进化分析中展现出卓越性能,但其在实际应用中仍存在若干有待深入探索的局限性.尽管ACCFV在HPV和禽流感病毒数据上表现良好,对于氨基酸变异率更高的病毒(如HIV或丙型肝炎病毒),当前特征提取策略的适应性仍显不足,需进一步优化以应对更高的序列多样性.未来研究工作将聚焦于ACCFV与深度学习等先进技术的融合,例如引入注意力机制以增强特征表示的可解释性,并拓展其在蛋白质功能预测、结构域识别等领域的应用能力,从而为大规模分子流行病学研究提供更强大且兼具可解释性的计算生物学工具.

参考文献:

[1] MCBRIDE A A. Human papillomaviruses: diversity, infection and host interactions [J]. Nature Reviews Microbiology, 2022, 20(2): 95-108.

[2] BRIANTI P, DE FLAMMINEIS E, MERCURI S R. Review of HPV-related diseases and cancers[J]. The New Microbiologica, 2017, 40(2): 80-85.

[3] AKBARI E, MILANI A, SEYEDINKHORASANI M, et al. HPV co-infections with other pathogens in cancer development: a comprehensive review [J/OL]. Journal of Medical Virology, 2023, 95(11)[2025-05-07]. https://doi.org/10.1002/jmv.29236.

[4] OYOUNI A A A. Human papillomavirus in cancer: infection, disease transmission, and progress in vaccines[J]. Journal of Infection and Public Health, 2023, 16(4): 626-631.

[5] HU Z, MA D. The precision prevention and therapy of HPV-related cervical cancer: new concepts and clinical implications [J]. Cancer Medicine, 2018, 7(10): 5217-5236.

[6] MUÑOZ N, BOSCH F X, DE SANJOSÉ S, et al. Epidemiologic classification of human papillomavirus types associated with cervical cancer[J]. The New England Journal of Medicine, 2003, 348(6): 518-527.

[7] GLENN W K, NGAN C C, AMOS T G, et al. High risk human *Papilloma* viruses (HPVs) are present in benign prostate tissues before development of HPV associated prostate cancer[J/OL]. Infectious Agents and Cancer, 2017, 12 [2025-05-07]. https://doi. org/10.1186/s13027-017-0157-2.

[8] YUSUPOV A, POPOVSKY D, MAHMOOD L, et al. The nonavalent vaccine: a review of high-risk HPVs and a plea to the CDC[J]. American Journal of Stem Cells, 2019, 8(3): 52-64.

[9] LAGHEDEN C, EKLUND C, LAMIN H, et al. Nationwide comprehensive human papillomavirus (HPV) genotyping of invasive cervical cancer[J]. British Journal of Cancer, 2018, 118(10): 1377-1381.

[10] XU X N, KONG R, LIU X Q, et al. Prediction of high-risk types of human papillomaviruses using reduced amino acid modes [J/OL]. Computational and Mathematical Methods in Medicine, 2020, 2020[2025-05-07]. https://doi.org/10.1155/2020/5325304.

[11] SILVA L LDA, TELES A M, SANTOS J M O, et al. Malignancy associated with low-risk HPV6 and HPV11: a systematic review and implications for cancer prevention[J/OL]. Cancers, 2023, 15(16)[2025-05-07]. https://doi.org/10.3390/cancers15164068.

[12] WOLF J, KIST L F, PEREIRA S B, et al. Human papillomavirus infection: epidemiology, biology, interactionshost, developmentcancer, prevention, and therapeutics[J/OL]. Reviews in Medical Virology, 2024, 34(3)[2025-05-07]. https://doi.org/10.1002/rmv.2537.

[13] SUDARSHAN S R, SCHLEGEL R, LIU X F. Two conserved amino acids differentiate the biology of high-risk and low-risk HPV E5 proteins [J]. Journal of Medical Virology, 2022, 94(9): 4565-4575.

[14] PAL A, KUNDU R. Human papillomavirus E6 and E7: the cervical cancer hallmarks and targets for therapy[J/OL]. Frontiers in Microbiology, 2020, 10[2025-05-07]. https://doi.org/10.3389/fmicb.2019.03116.

[15] ESTÊVÃO D, COSTA N R, GIL DA COSTA R M, et al. Hallmarks of HPV carcinogenesis: the role of E6, E7 and E5 oncoproteins in cellular malignancy[J]. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, 2019, 1862(2): 153-162.

[16] MÜNGER K, SCHEFFNER M, HUIBREGTSE J M, et al. Interactions of HPV E6 and E7 oncoproteins with tumour suppressor gene products[J]. Cancer Surveys, 1992, 12: 197-217.

[17] BAWONO P, DIJKSTRA M, PIROVANO W, et al. Multiple sequence alignment [M]//KEITH J M. Bioinformatics. New York: Humana New York, 2016: 167-189.

[18] EDGAR R C, BATZOGLOU S. Multiple sequence alignment[J]. Current Opinion in Structural Biology, 2006, 16(3): 368-373.

[19] KATOH K, MISAWA K, KUMA K I, et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform[J]. Nucleic Acids Research, 2002, 30(14): 3059-3066.

[20] POIROT O, O'TOOLE E, NOTREDAME C. Tcoffee@igs: a web server for computing, evaluating and combining multiple sequence alignments[J]. Nucleic Acids Research, 2003, 31(13): 3503-3506.

[21] MAGIS C, TALY J F, BUSSOTTI G, et al. T-coffee: tree-based consistency objective function for alignment evaluation [M]//RUSSELL D J. Multiple sequence alignment methods. Totowa: Humana Press, 2013: 117-129.

[22] ZHENG W, WUYUN Q, LI Y, et al. Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data[J]. Nature Methods, 2024, 21(2): 279-289.

[23] KAMINSKI K, LUDWICZAK J, PAWLICKI K, et al. pLM-BLAST: distant homology detection based on direct comparison of sequence representations from protein language models [J/OL]. Bioinformatics, 2023, 39 (10) [2025-05-07]. https://doi. org/10.1093/bioinformatics/btad579.

[24] BOHNSACK K S, KADEN M, ABEL J, et al. Alignment-free sequence comparison: a systematic survey from a machine learning perspective [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2023, 20(1): 119-135.

[25] STEINEGGER M, SÖDING J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets [J]. Nature Biotechnology, 2017, 35 (11): 1026-1028.

[26] ZIELEZINSKI A, VINGA S, ALMEIDA J, et al. Alignment-free sequence comparison: benefits, applications, and tools[J/OL]. Genome Biology, 2017, 18 (1)[2025-05-07]. https://doi. org/10.1186/s13059-017-1319-7.

[27] WALLACE I M, BLACKSHIELDS G, HIGGINS D G. Multiple sequence alignments [J]. Current Opinion in Structural Biology, 2005, 15(3): 261-266.

[28] TURANOV A A, LOBANOV A V, FOMENKO D E, et al. Genetic code supports targeted insertion of two amino acids by one codon [J]. Science, 2009, 323 (5911): 259-261.

[29] TREVINO V, OYERVIDES M, RAMÍREZ-CORREA G A, et al. Generating human papillomavirus (HPV) reference databases to maximize genomic mapping [J]. Archives of Virology, 2022, 167(1): 57-65.

[30] HE L, SUN S Y, ZHANG Q Y, et al. Alignment-free sequence comparison for virus genomes based on location correlation coefficient [J/OL]. Infection, Genetics and Evolution, 2021, 96 [2025-05-07]. https://doi. org/10.1016/j. meegid. 2021. 105106.

[31] TAMURA K, STECHER G, KUMAR S. MEGA11: molecular evolutionary genetics analysis version 11 [J]. Molecular Biology and Evolution, 2021, 38(7): 3022-3027.

[32] MUDAWI NAL, ALAZEB A. A model for predicting cervical cancer using machine learning algorithms [J/OL]. Sensors, 2022, 22 (11) [2025-05-07]. https://doi. org/10.3390/s22114132.

[33] SAITOU N, NEI M. The neighbor-joining method: a new method for reconstructing phylogenetic trees [J]. Molecular Biology and Evolution, 1987, 4(4): 406-425.

[34] EDGAR R C. MUSCLE: multiple sequence alignment with high accuracy and high throughput [J]. Nucleic Acids Research, 2004, 32(5): 1792-1797.

[35] THOMPSON J D, HIGGINS D G, GIBSON T J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice [J]. Nucleic Acids Research, 1994, 22(22): 4673-4680.

[36] ALTSCHUL S F, GISH W, MILLER W, et al. Basic local alignment search tool [J]. Journal of Molecular Biology, 1990, 215(3): 403-410.

[37] NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. BLAST+: command-line applications [DB/OL]. (2023-10-19)[2025-05-07]. https://ftp. ncbi. nlm. nih. gov/blast/executables/blast+/LATEST/.

[38] ARMSTRONG D J, ROMAN A. The relative ability of human papillomavirus type 6 and human papillomavirus type 16 E7 proteins to transactivate E2F-responsive elements is promoter- and cell-dependent[J]. Virology, 1997, 239(1): 238-246.

Evolutionary relationships and genotyping of HPV based on machine learning and amino acid position correlation coefficient method

HU Hualin, HU Lili, LIU Maoxing

(Beijing University of Civil Engineering and Architecture, Beijing 102616, China)

Abstract: In this study, a non-sequence-alignment method based on amino acid positional information, namely the amino acid correlation coefficient feature vector (ACCFV) method, was proposed for evolutionary analysis and genotyping of human papillomavirus (HPV). Traditional multiple sequence alignment (MSA) methods suffer from low computational

efficiency and high memory consumption when processing large-scale datasets. In contrast, the ACCFV method overcomes these limitations by constructing statistical measures of positional correlations between amino acids and converting amino acid sequences into numerical feature vectors. Amino acid sequences of eight HPV proteins (E6, E7, E1, E2, E4, E5, L1, and L2) were selected as target data. After feature extraction using ACCFV, a phylogenetic tree was constructed based on Euclidean distances between feature vectors, and four machine learning models were employed for classification prediction. The results showed that when the delay step size $L=1$, the ACCFV method achieved high consistency with the traditional MSA tool Muscle in evolutionary analysis, while significantly improving computational efficiency. Moreover, the Random Forest model achieved 100% classification accuracy. Compared to BLAST-Protein, ACCFV maintained 100% accuracy while substantially reducing processing time and required no batch operations. This study not only validates the feasibility and effectiveness of the ACCFV method in HPV research but also provides a novel technical approach for molecular epidemiological studies of other viruses.

Key words: HPV; amino acid sequence; machine learning; evolutionary analysis; subtype classification

附录A 去重后E6、E7、E1、E2、E4、E5、L1、L2验证集的预测准确率、F1-score与Recall值(表A1～A8)

附录B E6、E1、E2、E5、L1、L2蛋白在验证集上的预测准确率、F1-score和Recall值

(表B1～B6)

附录C 禽流感病毒亚型分类及结果