·监管技术·

基于Random Forest和UHPLC-QTOF-MS^{E}对不同来源龟甲基原的鉴定

王献瑞#，张佳婷#，张宇#，李明华，郭晓晗，荆文光，程显隆*，魏锋*（中国食品药品检定研究院，药品监管科学全国重点实验室，北京 102629）

摘要　目的：基于超高效液相色谱串联四极杆飞行时间质谱（UHPLC-QTOF-MS^E）分析并经数字量化处理，结合随机森林（Random Forest, RF）算法构建数据辨识模型，以实现中华草龟、巴西龟、台湾龟、鳄鱼龟、鳖甲基原的数字化鉴定。**方法：**经样品预处理后，对不同来源、不同批次的龟甲进行UPLC-QTOF-MS^{E}分析，并以混合样品为基准进行峰位校正、提取并经量化处理，获取反映多肽离子信息的精确质量数-保留时间数据对（Exact Mass Retention Time，EMRT）。然后基于信息增益率的特征筛选获取重要多肽离子信息，结合随机森林（RF）算法进行数据建模，同时基于内部交叉验证中的准确率（Acc）、精确率（P）、曲线下面积（AUC）等参数进行模型评价。最后基于最优模型进行龟甲基原的鉴定验证分析。**结果：**基于信息增益率的特征筛选，得到71个特征多肽信息，建立的RF模型具有优秀的辨识效果，准确率、精确率以及AUC均大于0.950且外部鉴定验证的正确率为100.0%。**结论：**基于UHPLC-QTOF-MS^{E}分析，并结合RF算法能够高效准确地实现不同来源龟甲基原的数字化鉴定，可为龟甲的质量控制及基原考证提供参考和帮助。

关键词：龟甲；基原鉴定；机器学习；随机森林；超高效液相色谱串联四极杆飞行时间质谱

中图分类号：R917　**文献标识码：**A　**文章编号：**1002-7777(2024)09-1008-012
doi:10.16153/j.1002-7777.20240512

Identification of Different Tortoiseshell's Species based on Random Forest and UHPLC-QTOF-MS^E

Wang Xianrui#, Zhang Jiating#, Zhang Yu#, Li Minghua, Guo Xiaohan, Jing Wenguang, Cheng Xianlong*, Wei Feng* (National Institutes for Food and Drug Control, State Key Laboratory of Drug Regulatory Science, Beijing 102629, China)

Abstract　Objective: Based on ultra-high performance liquid chromatography tandem quadrupole time-of-flight mass spectrometry (UHPLC-QTOF-MS^E) analysis and digital quantization, a data identification model was constructed by combining with the Random Forest (RF) algorithm to realize the digital identification of the species of Chinese tortoises, Brazilian tortoises, Taiwanese tortoises, alligator tortoises, and soft-shelled turtles.

基金项目：国家重点研发计划“中医药现代化”重点专项（编号 2023YFC3504105）；中国食品药品检定研究院学科带头人培养基金（编号 2023X10）
作者简介：王献瑞 Tel：（010）53851483；E-mail：niuyun006097@163.com
并列第一作者：张佳婷 Tel：（010）53851411；E-mail：18341441178@163.com
张宇 Tel：（010）53851411；E-mail：zhangyu55505@163.com
通信作者：程显隆 Tel：（010）53851475；E-mail：cxl@nifdc.org.cn
魏锋 Tel：（010）53852020；E-mail：weifeng@nifdc.org.cn

Methods: After sample pretreatment, different sources and batches of tortoiseshells were analyzed by UPLC-QTOF-MS^E. The peak positions were corrected, extracted, and quantified based on the mixed samples to obtain the data pairs of Exact Mass-Retention Time (EMRT) reflecting the information of peptide ions. Then the information about important peptide ions was obtained based on feature screening of information gain rate, combined with RF for data modeling. At the same time, the models were evaluated according to parameters such as accuracy (Acc), precision (P), and area under the curve (AUC) in internal cross-validation. Finally, the identification validation analysis of tortoiseshell species was carried out based on the optimal model. **Results:** Based on the feature screening of information gain rate, the 71 characteristic polypeptide information were obtained and the established RF model has excellent identification effect, with the accuracy, precision and AUC all greater than 0.950 and the correct rate of external identification validation was 100.0%. **Conclusion:** Based on the UHPLC-QTOF-MS^E analysis and combined with the RF algorithm, the digital identification of the species of the Tortoiseshell can be realized efficiently and accurately, which can provide reference and help for quality control and the species identification of the Tortoiseshell.

Keywords: tortoiseshell; species identification; machine learning; Random Forest; UHPLC-QTOF-MS^E

1 引言

龟甲始载于《神农本草经》，列为上品，作为一种传统中药，有着悠久的应用历史[1]。其具有滋阴潜阳、益肾强骨、养血补心、固经止崩之效，临床上常用于阴虚潮热、骨蒸盗汗、头晕目眩、崩漏经多等[2]。就其品种基原而言，《中华人民共和国药典》2020年版明确规定“本品为龟科动物乌龟*Chinemys reevesii*（*Gray*）的背甲及腹甲”[3]。然而，由于龟甲的品种基原单一、资源供不应求，在药材市场中经常有不法商家利用其他基原物种的龟甲伪品冒充正品，谋取非法利益，影响疗效和安全[4]。其中，较为常见的伪品有巴西龟、台湾龟、鳄鱼龟、鳖等其他来源的龟甲[5]。

为了实现不同来源龟甲品种基原鉴定及加强质量控制，众多研究人员开展了相关研究。钱敏等人基于UHPLC-QTOF-MS技术建立了专属特征肽鉴定方法，明确了乌龟、巴西龟、鳖的专属特征肽离子分别为$[M+2H]^+$=442.75、$[M+2H]^+$=400.24、$[M+2H]^+$=784.91[5]。徐清等人采用聚丙烯酰氨凝胶电泳（SDS-PAGE）技术对龟甲及其混伪品进行了鉴别，结果表明正品龟甲在14～19 kDa，有3条高丰度蛋白聚合在一起，可作为鉴别正品及伪品龟甲的初步依据[6]。基于线粒体基因组差异，Yang H和Li M等人开发了一种新的非测序方法来检测短DNA片段（约100 bp），用于快速鉴定龟甲和鳖甲[7-8]。上述研究均有助于龟甲的基原鉴定及质量控制。

随着中药数字化时代开启，机器学习中的随机森林（Random Forest，RF）算法已被广泛应用于中医药领域，比如基于数据建模的天然产物预测、疾病诊断、活性成分及靶点预测、基原鉴定、真伪鉴定等各方面[9-14]。同样，若能基于数字化表征及机器学习技术实现不同来源龟甲的鉴定分析，亦可进一步丰富龟甲的鉴定手段，有利于加强其质量控制。另一方面，龟甲作为典型的动物类中药，含有丰富的蛋白质、氨基酸及多肽。但从鉴别角度而言，氨基酸作为蛋白质的基本合成单元，不具有专属性。然而不同物种的DNA不同，由其指导合成的蛋白多肽序列往往会存在差异。因此，本研究从蛋白多肽角度获取不同基原物种龟甲的数字化表征，结合RF算法构建数据模型，以实现龟甲的基原鉴定，并探索其差异性多肽成分，以期丰富龟甲基原鉴定方法，加强质量控制。

2 材料与方法

2.1 仪器

美国Waters Acquity™ 超高效液相色谱仪，美国Waters Synapt G2 QTOF-MS质谱仪，FED-11S型号热循环干燥箱（德国BINDER公司），Climacell 111型号恒温恒湿培养箱（德国MMM公司），Milli-Q型号纯水净化仪（德国Merck Millipore有限公司）。

2.2 材料

本研究分析的龟甲来源包括6批鳄鱼龟甲、18

批巴西龟甲（常见伪品）、18批中华草龟甲（乌龟）、12批台湾龟甲（常见伪品）、7批鳖甲（常见伪品）[5-6]，共计61批次（其中50个批次用于数据建模，11个批次用于鉴定验证），所有样品经程显隆研究员鉴定，符合要求。龟甲样品的具体信息详见表1。

胰蛋白酶（质谱级）、亮氨酸脑啡肽（质谱级）、甲酸（色谱纯）均购自德国Sigma公司；乙腈（色谱纯）购自美国Fisher公司；碳酸氢铵（分析纯）购自国药集团化学试剂有限公司；纯净水为Milli-Q纯水。微孔滤膜（0.22 μm）购自德国Merck Millipore（Billerica，MA，USA）。

表 1 龟甲样品的具体信息

龟甲	拉丁名称	批号	基原
鳄鱼龟甲	*Chelydra serpentina* (*Linnaeus*)	EYG01~EYG06	鳄鱼龟
巴西龟甲	*Trachemys scripta elegans* (*Wied*)	BXG01~BXG18	巴西龟
乌龟甲	*Chinemys reevesii* (*Gray*)	ZHCG01~ZHCG18	中华草龟
台湾龟甲	*Ocadia sinensis* (*Gray*)	TWG01~TWG12	台湾草龟
鳖甲	*Trachemys scripta elegans* (*Wied*)	BJ01~BJ07	鳖

2.3 样品预处理

胰蛋白酶溶液配制：取质谱级胰蛋白酶，加入1%碳酸氢铵溶液溶解，制成1 mg·mL^{-1}胰蛋白酶溶液，用以酶解龟甲的多肽序列。

空白溶液配制：取1%碳酸氢铵溶液，用0.22 μm微孔滤膜过滤，取续滤液100 μL，置微量进样小瓶中，加入胰蛋白酶溶液10 μL，摇匀，在37 ℃条件下恒温酶解12 h。

供试品溶液配制：精密称取各批次龟甲50 g，加水500 mL，煮制8 h，过滤，滤液蒸干成胶状。然后取胶状粉末0.1 g，置50 mL量瓶中，加入1%碳酸氢铵溶液40 mL，超声处理（功率：500 W，频率：40 Hz）30 min，加入1%碳酸氢铵溶液稀释至刻度，摇匀，用0.22 μm微孔滤膜过滤，取续滤液100 μL，置微量进样瓶中，加胰蛋白酶溶液10 μL，摇匀，在37 ℃条件下恒温酶解12 h，制得各批次龟甲供试品溶液。此外，取所有批次样品混合后按上述方法制备混合质控样品用以峰位校正和提取。

2.4 UHPLC-QTOF-MS^{E}分析条件

液相条件：采用Waters Acquity^{TM} UPLC BEH C_{18}（2.1 mm×100 mm，1.7 μm）色谱柱，柱温为40 ℃，样品室温度为8 ℃，进样量为5 μL；流动相A为0.1%甲酸水溶液，流动相B为乙腈，流速为0.3 mL·min^{-1}，梯度洗脱程序详见表2。

质谱条件：离子化模式为ESI+，离子源温度为120 ℃；毛细管电压为3.0 kV，锥孔电压为20 V；除溶剂温度为450 ℃，除溶剂气体流速为600 L·h^{-1}；数据采集时采用亮氨酸脑啡肽（LE）进行实时质量校正，从而保证质谱数据采集的准确性与重复性。采用MS^E Continuum采集方式，扫描范围50~1200 Da，扫描时间为0.2 s；碰撞气体为高纯氩气，设定高、低2个能量通道交替采集质谱信息得到目标化合物的分子离子和碎片离子。低能量通道的碰撞能量为8 V，高能量通道的碰撞能量为10~40 V的梯度能量。

表 2 梯度洗脱程序

时间 /min	流动相 A/%	流动相 B/%
0.00	95	5
25.00	80	20
40.00	50	50
41.00	1	99
45.00	1	99
45.10	95	5
55.00	95	5

2.5 方法学考察

2.5.1 专属性考察

取空白溶液、不同基原龟甲供试品溶液进行检测分析，考察试验分析是否存在明显干扰。

2.5.2 精密度试验

取巴西龟甲（BXG09）供试品溶液，按“2.4”项下分析条件重复进样6次，记录酶解肽图。以选定的保留时间及其对应的质荷比为考察对象，计算RSD值，进行精密度考察。

2.5.3 重复性考察

取中华草龟甲（ZHCG14），按“2.3”项下方法平行制备5份供试品溶液，并按“2.4”项下分析条件进样，以随机选取的保留时间及其对应的质荷比为考察对象，计算RSD值，进行重复性考察。

2.5.4 稳定性考察

取鳄鱼龟甲（EYG03）供试品溶液，分别在0、2、6、8、12 h测定，以选取的保留时间及其对应峰的离子强度为考察对象，计算RSD值，进行稳定性考察。

2.6 质谱信息采集

经条件优化及方法学验证后，利用Waters Acquity™超高效液相系统及Waters Synapt G2 QTOF-MS系统，对供试品溶液进行LC-MS数据采集与分析，分别得到中华草龟、巴西龟、台湾龟、鳄鱼龟、鳖的基峰离子流图。

2.7 数字化特征筛选及数据建模分析

将不同批次龟甲的RAW文件导入Progenesis QI软件[15]，相关参数设定如下：ESI+，保留时间（Rt）：1~45 min，峰位校正及提取强度阈值设为不低于0.1%最高峰强度。以混合样品为校正标准，将LC-MS质谱中每个数据点转换成精确质量数-保留时间（Exact Mass Retention Time，EMRT）作为成分指标，离子强度为指标值的数据对，将50批次的龟甲量化数据导入Orange软件（2.3.5版本）用以数据建模。针对量化后的质谱数据，在Orange软件中，利用Rank功能中的信息增益率特征筛选算法进行筛选，并以筛选后的数据进行数据建模，以EMRT为变量指标，龟甲不同基原为判别目标，在Orange软件的Model模块中，选择RF算法建立辨识模型，并通过模型评估模块的准确度（Acc）、精确度（P）、曲线下面积（AUC）等指标进行模型评价，最后利用所建模型进行外部验证鉴定分析。

3 结果与讨论

3.1 方法学考察结果及讨论

3.1.1 专属性结果

专属性考察结果如图1所示。在“2.4”项分析条件下，不同批次、基原的龟甲均可被有效检测，空白溶液对供试品测定无明显干扰，结果表明专属性良好。

3.1.2 精密度考察结果

选取了巴西龟甲供试品8个色谱峰（A：568.7946，B：745.8721，C：449.7609，D：504.7789，E：789.9175，F：780.7051，G：620.3572，H：735.3932）进行提取，其保留时间和质荷比的RSD值均小于0.60%，详见表3和表4，结果表明仪器精密度良好。

表3 8个色谱峰保留时间在精密度试验考察中的结果

进样次数	保留时间 /min							
	A	B	C	D	E	F	G	H
1	3.02	7.05	9.02	12.89	14.57	16.42	21.46	23.93
2	3.01	7.03	9.03	12.92	14.59	16.42	21.43	23.93
3	3.01	7.00	8.96	12.87	14.54	16.39	21.41	23.90
4	3.01	7.03	9.02	12.88	14.56	16.38	21.46	23.93
5	2.98	7.03	9.05	12.92	14.60	16.43	21.44	23.94
6	2.99	7.03	9.03	12.90	14.59	16.43	21.46	23.93
平均值 /min	3.00	7.03	9.02	12.90	14.58	16.41	21.44	23.93
RSD/%	0.50	0.23	0.34	0.16	0.15	0.13	0.10	0.06

表 4 8 个色谱峰质荷比在精密度试验考察中的结果

进样次数	质荷比（*m/z*）							
	A	B	C	D	E	F	G	H
1	568.7946	745.8721	449.7609	504.7789	789.9175	780.7051	620.3572	735.3932
2	568.7950	745.8729	449.7606	504.7796	789.9164	780.7047	620.3580	735.3929
3	568.7950	745.8729	449.7607	504.7790	789.9175	780.7046	620.3574	735.3922
4	568.7956	745.8716	449.7610	504.7786	789.9174	780.7048	620.3577	735.3932
5	568.7946	745.8733	449.7609	504.7783	789.9171	780.7052	620.3568	735.3921
6	568.7953	745.8729	449.7605	504.7792	789.9171	780.7054	620.3575	735.3918
平均值	568.7950	745.8726	449.7608	504.7789	789.9172	780.7050	620.3574	735.3926
RSD/%	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

3.1.3 重复性考察结果

选取中华草龟供试品8个色谱峰（A：745.8716, B：449.7603, C：776.8912, D：794.8913, E：795.9104, F：727.3761, G：971.4752, H：991.5135）进行提取，记录上述8个色谱峰（A、B、C、D、E、F、G、H）的保留时间和质荷比并计算RSD值。其保留时间和离子强度RSD值均小于0.30%，详见表5和表6，结果说明方法的重复性良好，符合要求。

表 5 8 个色谱峰保留时间在重复性试验考察中的结果

样品数	保留时间 /min							
	A	B	C	D	E	F	G	H
1	7.20	9.26	10.43	12.85	14.16	20.02	25.74	28.24
2	7.24	9.26	10.45	12.87	14.16	20.02	25.74	28.24
3	7.23	9.28	10.46	12.86	14.17	20.03	25.77	28.24
4	7.20	9.23	10.45	12.86	14.22	20.01	25.76	28.26
5	7.21	9.25	10.43	12.87	14.17	20.03	25.74	28.24
6	7.24	9.29	10.45	12.86	14.17	20.04	25.76	28.24
平均值 /min	7.22	9.26	10.45	12.86	14.18	20.03	25.75	28.24
RSD/%	0.26	0.23	0.12	0.06	0.16	0.05	0.05	0.03

表 6 8 个色谱峰质荷比在重复性试验考察中的结果

样品数	质荷比（*m/z*）							
	A	B	C	D	E	F	G	H
1	745.8716	449.7603	776.8912	794.8913	795.9104	727.3761	971.4752	991.5135
2	745.8723	449.7604	776.8901	794.8907	795.9111	727.3762	971.4744	991.5136
3	745.8728	449.7603	776.8907	794.8900	795.9109	727.3762	971.4743	991.5129
4	745.8720	449.7603	776.8918	794.8915	795.9108	727.3765	971.4752	991.5132
5	745.8718	449.7601	776.8911	794.8914	795.9106	727.3755	971.4747	991.5130
6	745.8724	449.7601	776.8909	794.8914	795.9110	727.3761	971.4747	991.5127
平均值	745.8722	449.7603	776.8910	794.8911	795.9108	727.3761	971.4748	991.5132
RSD/%	0.0001	0.0000	0.0001	0.0001	0.0000	0.0000	0.0000	0.0000

3.1.4 稳定性考察结果

选取了8个色谱峰（A：745.8725，B：449.7600，C：697.3407，D：771.8804，E：892.9515，F：727.3770，G：981.5090，H：1064.5565）进行提取，上述8个色谱峰（A、B、C、D、E、F、G、H）在12 h内均可检出。记录上述8个色谱峰的保留时间以及离子强度，计算RSD值。其保留时间以及离子强度的RSD值均小于3.00%，详见表7和表8，结果说明样品稳定性良好。

表 7 8 个色谱峰保留时间在稳定性试验考察中的结果

进样时间	保留时间 /min							
	A	B	C	D	E	F	G	H
0 h	7.09	9.11	9.50	10.89	13.24	19.87	28.84	31.87
2 h	7.11	9.14	9.51	10.91	13.27	19.86	28.86	31.88
6 h	7.11	9.11	9.50	10.88	13.25	19.87	28.86	31.88
8 h	7.09	9.11	9.50	10.89	13.22	19.89	28.86	31.89
12 h	7.12	9.08	9.51	10.89	13.27	19.90	28.85	31.88
平均值 /min	7.10	9.11	9.50	10.89	13.25	19.88	28.85	31.88
RSD/%	0.19	0.23	0.06	0.10	0.16	0.08	0.03	0.02

表 8 8 个色谱峰离子强度在稳定性试验考察中的结果

进样时间	离子强度（I）							
	A	B	C	D	E	F	G	H
0 h	898249	2741230	1322880	1295780	1818500	6643020	15997400	2373120
2 h	914343	2749940	1323710	1292140	1928350	6906220	15311300	2301340
4 h	877763	2642990	1270690	1284380	1828550	6986910	15310800	2273150
8 h	893475	2661390	1286430	1213270	1842490	6722830	15560400	2299910
12 h	918228	2608700	1298410	1297820	1804080	6728760	15879300	2351440
平均值	900411.6	2680850	1300424	1276678	1844394	6797548	15611840	2319792
RSD/%	1.82	2.32	1.77	2.81	2.66	2.10	2.04	1.77

在进行方法学考察时，参考了《分析方法验证指导原则》以及《生物样品定量分析方法验证指导原则》，着重从专属性、精密度、重复性、稳定性几方面进行考察，结果表明试验方法专属性良好、仪器精密度和重复性良好，能够满足分析需要。在方法学考察时，以保留时间8 min、20 min为划分，每个区段选取不少于1个离子信息（保留时间–质荷比或者保留时间–离子强度）进行记录（每个区段内则是随机选取），总共选取8个，并计算其RSD值。方法学考察有利于评估在不同时空采集数据的一致性。然而并未进行准确度考察，主要是因为LE作为LockSpray可实时进行质量校正，从而保证检测结果的准确性。

3.2 质谱信息采集结果及讨论

基于“2.3”项下样品预处理和“2.4”项下试验分析条件，在不同时间进行采集，每次采集周期为12 h，获取了61个批次龟甲样品的质谱信息，空白、不同基原龟甲的基峰离子流图（专属性）详见图1。

如图1所示，从整体而言，不同基原龟甲的基峰离子流图存在差异，表明其酶解后的多肽序列确实不同，而这种差异为基于数据建模开展数字化非靶向鉴定奠定了基础。当对少量样品进行分析时，仅仅通过谱图比对，就可以实现中药材的鉴定。但随着样品的增加及个体差异的凸显，仅仅通过人工谱图比较，难免效率降低，且基于余弦值或欧式距离等的谱图相似度比较，更多强调谱图轮廓、方向的一致性，精准度降低。因此，有必要进一步探索龟甲基原的数字化鉴定。

在UHPLC–QTOF–MS^{E}分析中，样品预处理方法参考了《中华人民共和国药典》2020年版四部通则3405肽图检查法中的溶剂条件，同时参考了阿胶【鉴别】项下的酶解条件，最终确定了酶的用量为供试品续滤液100 μL中加胰蛋白酶溶液10 μL以及酶解时间12 h[16–17]。而且比较了碰撞能分别为10～30、10～40和10～50 V时，各批次龟甲的质谱信息量，结果表明当碰撞能为10～40 V时，质谱信息最为丰富。

3.3 数字化特征筛选结果与讨论

在上述特征筛选过程中，采用了基于信息增益率的特征筛选算法[18]。该方法是基于信息增益衍生而来。信息增益算法会倾向于选择取值较多的特征变量，但引入该特征变量熵增比值，则会修正这一问题，从而避免筛选不同基原龟甲的共有多肽离子，聚焦于不同基原龟甲的特异性多肽离子，既可以满足删减冗杂数据进行数据建模，又有助于探索差异性多肽离子。因此，最终选择该算法进行特征筛选。

以混合样品为参比，对50个批次不同品种的龟甲质谱信息进行数据转化，从而保证数据完整。最后共得到1623个EMRT数据对，其中包括不同基原龟甲的特异性多肽离子及共有的多肽离子数据。将上述数据导入Orange软件（2.3.5版本），通过Rank功能中的信息增益率特征筛选算法，以大于

0.6为筛选标准，筛选得到对于区分不同品种龟甲重要的71个特征变量（EMRT数据对），具体详见表9。

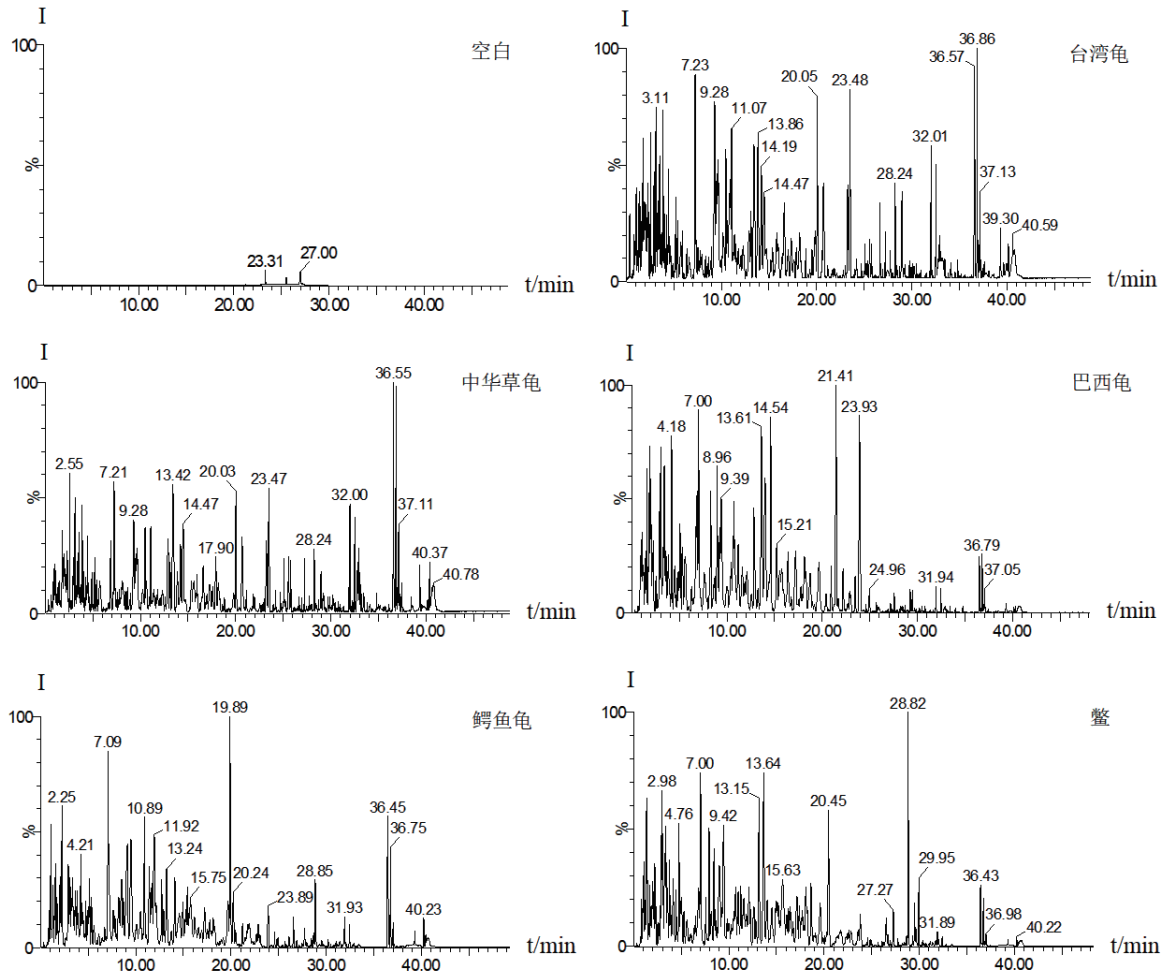

图 1　空白及不同基原龟甲的基峰离子流图（专属性）

表 9　71 个重要特征变量及其信息增益率值

特征变量	增益率值	特征变量	增益率值	特征变量	增益率值
14.63min_884.9539*m/z*	0.835	15.54min_758.3643*m/z*	0.673	1.31min_450.7314*m/z*	0.628
1.03min_465.4207*m/z*	0.813	17.84min_1011.7288*m/z*	0.670	13.47min_936.9782*m/z*	0.628
15.02min_884.9540*m/z*	0.764	17.84min_1011.9798*m/z*	0.669	1.39min_171.1134*m/z*	0.627
4.37min_673.9952*m/z*	0.755	15.10min_962.4753*m/z*	0.667	4.43min_621.8022*m/z*	0.627
11.78min_958.4691*m/z*	0.755	8.46min_278.6646*m/z*	0.667	16.00min_855.7395*m/z*	0.625
11.78min_958.8033*m/z*	0.755	8.80min_734.6900*m/z*	0.665	9.66min_615.3233*m/z*	0.625
20.27min_838.7611*m/z*	0.755	9.29min_579.9566*m/z*	0.659	4.91min_197.1292*m/z*	0.622

续表 9

特征变量	增益率值	特征变量	增益率值	特征变量	增益率值
28.82min_981.8430*m/z*	0.755	9.29min_580.6249*m/z*	0.659	4.20min_523.5595*m/z*	0.620
9.65min_932.4413*m/z*	0.753	4.37min_673.6611*m/z*	0.658	9.06min_425.2096*m/z*	0.619
28.65min_725.4564*m/z*	0.736	28.82min_981.5086*m/z*	0.658	7.75min_869.9058*m/z*	0.618
12.73min_578.6248*m/z*	0.732	5.37min_730.3726*m/z*	0.656	12.38min_580.9616*m/z*	0.614
15.98min_1282.6057*m/z*	0.732	14.04min_593.2984*m/z*	0.652	20.47min_795.9348*m/z*	0.613
13.32min_798.9033*m/z*	0.731	7.67min_1088.0208*m/z*	0.650	8.43min_769.8832*m/z*	0.610
2.10min_523.7603*m/z*	0.723	9.70min_646.8397*m/z*	0.650	7.72min_530.2933*m/z*	0.605
11.20min_575.2924*m/z*	0.714	20.27min_838.0940*m/z*	0.644	12.38min_870.9383*m/z*	0.604
1.12min_328.1748*m/z*	0.712	15.74min_629.9932*m/z*	0.643	1.73min_448.2153*m/z*	0.604
1.03min_581.5244*m/z*	0.710	12.06min_754.4089*m/z*	0.641	10.09min_883.1003*m/z*	0.604
14.68min_534.9338*m/z*	0.706	11.47min_734.3618*m/z*	0.639	11.47min_1100.5366*m/z*	0.604
21.70min_1047.7575*m/z*	0.705	22.93min_1043.7592*m/z*	0.634	4.22min_641.3137*m/z*	0.603
11.00min_771.8801*m/z*	0.704	15.60min_849.4041*m/z*	0.630	12.46min_750.3613*m/z*	0.603
15.74min_629.6597*m/z*	0.696	8.46min_425.2083*m/z*	0.629	8.64min_816.4094*m/z*	0.602
9.66min_250.1537*m/z*	0.692	11.74min_574.6247*m/z*	0.629	29.98min_781.5181*m/z*	0.602
11.20min_574.6247*m/z*	0.690	23.97min_650.3390*m/z*	0.628	35.01min_385.1910*m/z*	0.602
9.88min_937.1181*m/z*	0.689	15.75min_943.9855*m/z*	—	—	—

总体来讲，上述71个特征变量（EMRT数据对）是以混合样品为参考，经过峰位校正、提取以及信息增益率算法结合筛选后得到的特征变量，其中共有多肽离子表征数据以及特异性多肽离子表征数据，均可在不同批次、基原的龟甲样品中有稳定的呈现，虽有数据偏差的波动，但通过合理偏差设定（保留时间偏差≤0.20 min；$\Delta m/z$≤10 ppm）即可稳定控制，且用于外部验证在不同时期采集的不同批次的龟甲样品经数据转化后依然可以得到上述71个特征变量，因此，经算法筛选的71个特征变量可在不同样品中稳定存在，有利于基于以上71个特征变量实现龟甲基原鉴定分析。

3.4 数据建模结果与讨论

基于“3.3”项筛选得到的71个特征数据变量，在Orange软件的Model模块中，选择Random Forest算法构建数据模型，同时以2/3的数据信息作为训练集，1/3的数据信息作为测试集，分别进行10、20、50次内部验证，其曲线下面积、准确度、平衡分数、精确度、查全率等评价参数详见表10。

表 10 基于 71 个特征变量所建 RF 模型评价参数

模型	曲线下面积	准确率	平衡分数	精确度	查全率	内部验证次数
RF	0.999	0.965	0.964	0.967	0.953	10
RF	0.998	0.965	0.965	0.965	0.967	20
RF	0.998	0.971	0.971	0.971	0.971	50

曲线下面积是用来衡量分类模型整体效果的一个量化指标；准确率表示模型正确预测的样本数占总样本数的比例；查全率表示模型正确预测的正样本数占所有真正样本数的比例；平衡分数是精确度和召回率的调和平均值，用于平衡精确度和召回率；精确度代表的是模型预测为正样本中实际也为正样本的所占比例；上述参数越接近于1.000，则表明模型效果越好。

如表10所示，无论进行10、20、50次交叉验证，RF数据模型的曲线下面积均大于0.990，表明RF模型效果显著；其准确率均不低于0.960，表明模型准确率满足鉴定需求[19]；精确度不低于0.965，亦体现了模型分类效果较佳[20]。此外，平衡分数及查全率也均大于0.950。综上所述，基于筛选的71个特征变量（多肽离子）所构建的RF模型对于龟甲不同基原具有良好的鉴定辨识效果，准确率较高。

在RF模型构建中，进行了超参数优化，发现当随机森林树数量为10、节点分裂考虑变量数为5、叶最大深度为3时，模型效果最优。此外，在相同条件下，还比较了利用筛选前1623个特征变量和筛选后71个特征变量所构建的RF模型效果。基于1623个特征变量构建的RF模型的评估参数如表11所示。

表 11 基于 1623 个特征变量所建 RF 模型的内部交叉验证评价参数

模型	曲线下面积	准确率	平衡分数	精确度	查全率	内部验证次数
RF	0.988	0.906	0.906	0.911	0.906	10
RF	0.992	0.915	0.915	0.919	0.915	20
RF	0.991	0.914	0.914	0.917	0.914	50

对比表10和表11可以发现，利用特征筛选后，71个特征变量构建的RF模型的评价参数有显著提高，比如在相同条件下，准确率至少提高了0.050，精确度至少提高了0.046等。由此可见，在1623个数据变量中确实存在冗杂、不利于龟甲基原鉴定的“无效数据”，通过特征筛选可有效剔除该部分数据，从而提高鉴定辨识的准确率和精确度，亦说明了特征筛选的必要性。

3.5 基于RF模型的鉴定验证及差异性多肽离子初步探索

对基于71个特征变量所构建的RF模型进行外部验证，首先利用未用于建模的样本数据，经特征筛选得到的71个[Rt–*m*/*z*–I]作为特征变量进行输入，通过RF模型对11批龟甲样品进行基原数字化鉴定，结果详见表12。如表12所示，经过外部验证，11批样品（4批巴西龟甲，3批中华草龟甲，2批台湾龟甲，1批鳖甲和1批鳄鱼龟甲）在所构建的RF模型中均可被正确鉴定识别，正确率100.000%，与实际情况相符，这表明基于多肽离子量化数据和RF算法的数据模型具有一定的实用价值，能够有效实现龟甲不同基原的鉴别。该方法对于龟甲及其中药不同基原的鉴定分析具有重要意义。

表 12 基于 RF 模型的龟甲基原鉴定结果

龟甲	模型鉴定结果	龟甲	模型鉴定结果
巴西龟	巴西龟	中华草龟	中华草龟
巴西龟	巴西龟	鳄鱼龟	鳄鱼龟
巴西龟	巴西龟	台湾龟	台湾龟
巴西龟	巴西龟	台湾龟	台湾龟
中华草龟	中华草龟	鳖甲	鳖甲
中华草龟	中华草龟	—	—

在此基础上，根据“3.3”项信息增益率的特征筛选，还初步探索了不同基原龟甲的差异性多肽离子。信息增益率越大，该变量对于区分龟甲的不同基原贡献越大，结合信息增益率的特征筛选偏向于取值较少的变量。因此，基于信息增益率的特征筛选，进一步从筛选后的71个多肽数据变量中得到了一些特异性多肽离子，比如双电荷多肽9.29 min_579.9566 *m/z*仅能在巴西龟甲中被检测到，离子强度不低于1.7×10^6，而在其他龟甲中基本处于基线水平；双电荷多肽11.78 min_958.8033 *m/z*及9.65 min_932.4413 *m/z*则仅能在鳄鱼龟甲中被检测到；双电荷多肽20.47 min_795.9348 m/z仅能在台湾龟甲中被检出；鳖甲的专属性多肽离子有单电荷多肽28.65 min_725.4564 *m/z*等。对于中华草龟而言，未能找到其专属性多肽离子，但双电荷离子13.47 min_936.9782 *m/z*在中华草龟甲中的离子强度是其在鳖甲中离子强度的20倍以上，而在其他品种的龟甲中未检出。因此，13.47 min_936.9782 *m/z*也可被视为中华草龟的辨识性多肽离子。

3.6 研究优势、局限及展望

本研究创新之处在于将随机森林机器学习算法与多肽离子量化表征相结合，开展龟甲基原鉴定分析。通过随机森林算法构建鉴定模型即可实现龟甲基原的数字化鉴定分析，在一定程度上可避免人为主观因素的干扰，且无需鉴定多肽的氨基酸序列。另一方面，通过筛选得到的71个多肽特征变量，相较于单个或少数几个多肽离子的专属性更强、精确度更高，因此，该方法可大大提高鉴定分析的效率以及龟甲基原鉴定的精确度。然而，本研究依然存在局限：研究中分析的样品共计61批，样品数量较少，后续研究需要增大样本量，并基于筛选得到的71个多肽特征变量进行龟甲基原的数字化鉴定分析。而且本研究针对中华草龟甲、巴西龟甲、台湾龟甲、鳄鱼龟甲、鳖甲进行了鉴定，虽然包括常见的龟甲伪品—鳖甲、巴西龟甲、台湾龟甲等，但龟甲种类繁多，难以全部囊括，后续亦将扩大品种进一步探索。此外，基于信息增益率特征筛选算法，虽初步探索了不同龟甲的特异性多肽离子，但并未确定多肽的氨基酸序列组成。后续还将针对初步筛选的差异性多肽离子，进一步优化质谱条件，探索特征性离子对及龟甲近缘种的差异性指标，用于龟甲药材或含龟甲中成药的鉴定。

4 结论

本研究基于UHPLC-QTOF-MS^E技术对不同基原龟甲多肽进行分析，并经量化处理后进行特征筛选得到71个特征变量数据，进一步结合RF算法构建了鉴定模型并用于外部验证，结果表明基于UHPLC-QTOF-MS^E分析的多肽量化数据和RF算法模型能够有效实现龟甲不同基原的鉴定，该方法可为龟甲的质量控制及基原考证提供参考和帮助，亦在一定程度上有助于实现中药的数字化鉴定。

参考文献：

[1] 黄清杰，张中华，徐志伟. 龟甲药用历史与研究进展[J]. 湖北中医杂志，2021，43（5）：64-66.

[2] 刘俐，何清湖，唐宇，等. 龟甲的现代研究进展[J]. 湖南中医杂志，2020，36（7）：181-183.

[3] 中华人民共和国药典：一部[S]. 2020：187-188.

[4] 张琳，戴仕林，党静洁，等. 不同品种龟甲的光谱学特征研究[J]. 现代中药研究与实践，2023，37（4）：

5–10.

[5] 钱敏，刘宇文，刘晓华，等. 基于UPLC–QTOF–MS技术鉴别10种龟鳖甲类中药的真伪[J]. 中华中医药杂志，2022，37（6）：3434–3440.

[6] 徐清，李梦，罗雪梅，等. 采用聚丙烯酰氨凝胶电泳技术鉴别龟甲及其混伪品[J]. 中国现代中药，2019，21（9）：1251–1255.

[7] Li M, Wang M, Zhou Y, et al. Identification and Characteristics of Testudinis Carapax et Plustrum based on Fingerprint Profiles of Mitochondrial DNA Constructed by Species–specific PCR and Random Amplified Polymorphic DNA[J]. Mitochondrial DNA B Resour，2018，3（2）：1009–1012.

[8] Yang H, Yu P, Lu Y, et al. A Novel Non–sequencing Approach for Rapid Authentication of Testudinis Carapax et Plastrum and Trionycis Carapax by Species–specific Primers[J]. R Soc Open Sci，2018，5（4）：172140.

[9] Xie Q, Cui M, Wu ZD, et al. Traditional Chinese Medicine Information Digitalization Discussion[J]. J Altern Complement Med，2010，16（11）：1207–1209.

[10] Wang Y, Shi X, Li L, et al. The Impact of Artificial Intelligence on Traditional Chinese Medicine[J]. Am J Chin Med，2021，49（6）：1297–1314.

[11] Ma S, Liu J, Li W, et al. Machine Learning in TCM with Natural Products and Molecules：Current Status and Future Perspectives[J]. Chin Med，2023，18（1）：43.

[12] Yuan L, Yang L, Zhang S, et al. Development of a Tongue Image–based Machine Learning Tool for the Diagnosis of Gastric Cancer： A Prospective Multicentre Clinical Cohort Study[J]. E Clinical Medicine，2023，57：101834.

[13] Chen H, He Y. Machine Learning Approaches in Traditional Chinese Medicine：A Systematic Review[J]. Am J Chin Med，2022，50（1）：91–131.

[14] Dong X, Zheng Y, Shu Z, et al. TCMPR： TCM Prescription Recommendation based on Subnetwork Term Mapping and Deep Learning[J]. Biomed Res Int，2022，2022：4845726.

[15] Zhang J, Yang W, Li S, et al. An Intelligentized Strategy for Endogenous Small Molecules Characterization and Quality Evaluation of Earthworm from Two Geographic Origins by Ultra–high Performance HILIC/QTOF MS（E）and Progenesis QI[J]. Anal Bioanal Chem，2016，408（14）：3881–3890.

[16] 中华人民共和国药典：一部[S]. 2020：197–198.

[17] 中华人民共和国药典：四部[S]. 2020：317–319.

[18] 朱燚丹, 陈兴荣, 李秋萍. 基于信息增益率的超高维变量选择[J]. 统计与决策, 2021, 37（22）：18–21.

[19] Jiang T, Gradus JL, Rosellini AJ. Supervised Machine Learning： A Brief Primer[J]. Behav Ther，2020，51（5）：675–687.

[20] Rainio O, Teuho J, Kl é n R. Evaluation Metrics and Statistical Tests for Machine Learning[J]. Sci Rep，2024，14（1）：6086.

（收稿日期 2024年6月28日 编辑 王雅雯）