

基于氨基酸性质的二肽稳定性定量预测模型研究^{*}

王远强, 郭海琼, 王昱璇, 张玉萍, 张 娅, 林治华

(重庆理工大学 药物化学与分子药理学重庆市重点实验室 靶向药物筛选与活性评价重庆市重点实验室, 重庆 40005)

摘要:【目的】二肽具有生物活性高、易于合成等优点,但在机体内有较差的代谢稳定性,易被降解,故基于氨基酸性质,对二肽稳定性进行定量预测研究,为设计稳定性好的二肽分子提供理论依据。【方法】基于 209 个二肽分子在不同时间段的降解率,使用偏最小二乘法,将逐步回归筛选变量与支持向量机、随机森林、多元线性回归等方法相结合,建立多肽降解率与氨基酸理化性质之间的定量预测模型。【结果】最为显著的是对二肽 60 min 降解率所建模型,对训练集和测试集分别具有良好的估计能力($R^2 > 0.68$, $Q^2 > 0.57$)与预测能力($R^2 > 0.54$),能够有效预测二肽分子的降解率;而且基于多元线性回归系数计算的氨基酸贡献能够发现影响二肽稳定性的重要氨基酸,可以指导高稳定性二肽分子的合理设计。【结论】建立的预测模型方法简单,物理意义明确,多种方法均能获得较为理想的预测模型,确保了预测结果的准确性,可用于指导设计和筛选稳定性好的二肽分子。

关键词:二肽;稳定性;降解率;定量预测模型;合理设计

中图分类号:R392.5

文献标志码:A

文章编号:1672-6693(2018)06-0117-07

到目前为止,在生物体内发现的多肽已达数万种^[1]。在药用方面,它们有广泛的适应症、较高的安全性以及明显的疗效。目前,在肿瘤、肝炎、糖尿病、艾滋病等疾病的预防、诊断和医治领域,多肽药物已被广泛应用^[2]。目前,多肽药物在心血管疾病治疗方面的研究越来越多^[3]。活性多肽大部分属于小分子物质,即小分子肽^[4],属于心血管自稳态调节的重要部分。在高血压发病过程中,这些活性多肽对高血压的调节及机制的研究起着重要作用,目前在这方面的基础研究及临床研究均已取得很大进展^[5]。

多肽类药物代谢机制研究表明,这类物质的消除途径与内源性或膳食蛋白质的分解代谢途径相同,即主要经过受体内蛋白酶作用而被水解代谢^[6]。尽管多肽药物有很高的药用利用价值,但是存在着不可忽视的问题或缺点制约着它的应用发展,原因在于多肽药物在机体内代谢稳定性较差、易迅速地被降解、消除半衰期较短、生物利用度低且不能口服等。生物有效性在药物活性成分开发中具有重要作用,如果将多肽类直接作为药物开发,将有很大局限性^[7]。多肽类物质稳定性差主要是基于肽的特殊分子结构,以及主链氨基酸降解和侧链氨基酸残基结构的变化。因此,对肽类药物自身的分子结构进行研究和改造,将对改善多肽的稳定性并使之发挥较高的体内治疗效果至关重要。

多肽的定量构效关系(QSAR)研究主要表现在两个方面:对多肽进行结构表征和构建 QSAR 数学模型。在多肽的各种 QSAR 模型中,普遍采用氨基酸的结构描述参数对多肽结构进行表征,定量描述多肽化学结构和氨基酸残基。在建立的 QSAR 模型中,常将线性回归、偏最小二乘法(PLS)、遗传算法(GA)和人工神经网络(ANN)等统计方法相结合,来建立化合物的定量结构参数与生物活性之间的函数关系^[8]。多肽结构表征与多种 QSAR 建模方法对多肽结构设计和稳定性的研究,已被普遍用于医学和毒理学领域,并且已被证明可预测苦味肽的活性。另有将生物活性定量构效的方法,应用于分析蚕蛹酶解蛋白制备血管紧张素转化酶抑制肽的可能性研究^[9]。

本研究采用氨基酸物理化学性质表征多肽结构,基于 209 个二肽分子在不同时间段的降解率数据集,使用

^{*} 收稿日期:2018-05-22 修回日期:2018-10-10 网络出版时间:2018-10-25 10:41

资助项目:国家自然科学基金(No. 31400667);重庆市教委科学技术研究项目(No. KJ1400932, No. KJ1500902, No. KJ160008);重庆市基础研究
研究与前沿技术研究项目(No. cstc2014jcyjA10044);重庆高校创新团队建设计划项目(No. CXTDX201601031)

第一作者简介:王远强,男,博士,副教授,研究方向为生物信息学和药物化学, E-mail: wangyqnn@cqut.edu.cn;通信作者:张娅,女,讲师,
E-mail: zhang_ya@cqut.edu.cn

网络出版地址: <http://kns.cnki.net/kcms/detail/50.1165.N.20181025.1041.012.html>

PLS,将逐步回归(SWR)筛选变量与支持向量机(SVM)、随机森林(RF)、多元线性回归(MLR)等方法相结合,建立多肽降解率与氨基酸理化性质之间的定量预测模型,为多肽的稳定性研究提供理论指导。

1 数据与方法

1.1 数据集

本研究中的 209 个二肽分子来自文献[10],文献中报道了二肽在 5,30,60 min 的降解率。本研究分别对不同时间的降解率建立定量预测模型。在数据处理中,二肽降解率超过 100.0%的按 100.0%处理、低于 0.0%的按 0.0%处理。在建模过程中,将数据集按降解率从高到低排序,随机选择 25%(53 个)的样本作为外部测试集,用于检测模型的可靠性;剩余的 75%(156 个)作为训练集,用于建立定量预测模型。

1.2 结构表征

用于多肽结构表征的氨基酸物理化学性质来自 AAindex 数据库^[11],多肽的每个氨基酸用 195 种性质(包括 149 种疏水性性质和 46 种极性性质),则每个二肽由 $2 \times 195 = 390$ 个变量构成的向量表示。

1.3 模型构建

由于每个二肽由 390 个变量表征,为了避免多重贡献性,使用 SWR,从初始变量中筛选重要参数^[12],SWR 引入(F_1)与剔除(F_2)变量的阈值分别为 0.05 和 0.1,并用变量膨胀因子(VIF)检验所选变量的多重共线性^[13],然后基于重要参数分别使用 MLR^[14]、SVM^[15]和 RF^[16]构建模型。为了进一步对比模型的可靠性,分别使用 PLS^[17]、SVM^[18]和 RF^[19]对初始变量的数据集构建定量预测模型。本研究建模工作均用 Matlab 软件完成。

2 结果与讨论

2.1 对 5 min 降解率的建模结果分析

针对二肽 5 min 的降解率数据集,逐步回归从 390 个变量中筛选出 7 个重要变量,其中 4 个变量用于表征第一位的氨基酸、3 个变量用于表征第二位的氨基酸,变量膨胀因子均小于 4,低于临界值 5(图 1a),表明所选择变量的自相关性较小,能够有效避免定量预测模型的过拟合,所用建模变量详见表 1。

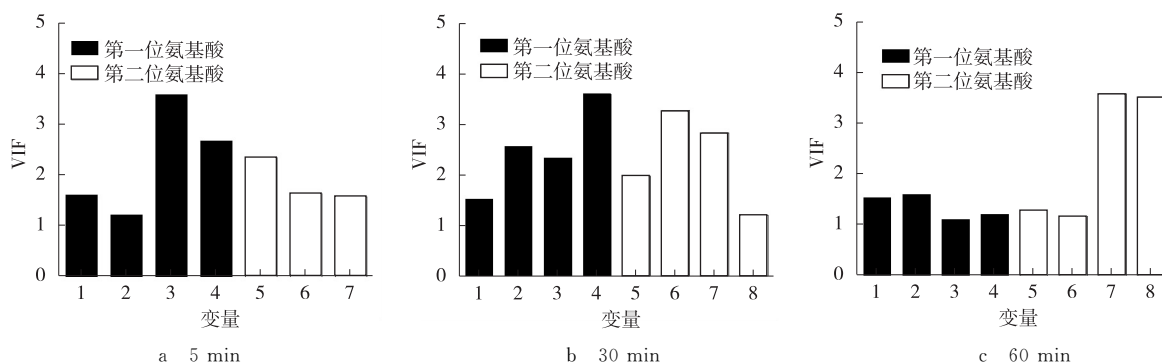


图 1 基于 QSAR 模型在不同时间下氨基酸对二肽降解的贡献

Fig. 1 The contribution of amino acids for degradation of dipeptide based on QSAR model for different time

本研究分别基于 SWR 筛选的重要变量,使用 SVM,RF,MLR 等方法建立定量预测模型;利用 PLS,SVM,RF 方法的多变量处理优势,对所有变量直接构建定量预测模型,其中用 PLS 方法最终选择出 7 个主成分建立模型,以便与其他的建模方法进行对比,建模结构详见表 1。

对比表 1 中不同方法所建定量预测模型发现,除 SWR-SVM 方法所建定量预测模型预测不理想外,其余方法建立的预测模型均具有良好的稳定性($R^2 > 0.50$, $Q^2 > 0.31$)与外部预测能力($R^2 > 0.40$),其中 PLS 和 SWR-RF 方法所建模型的稳定性($R^2 > 0.6$)和预测能力($R^2 > 0.5$)优于其他几种方法;尽管 SWR-MLR 方法所建模型的稳定性最好($Q^2 = 0.455$),但是外部预测能力($R^2 = 0.406$)还有待提高。降解率观测值与计算值的相关性图(图 2)显示,PLS,RF 及 SWR-RF 方法所建模型对训练集和测试集的预测值与实验观测值相关性均较好,而 SVM 所建模型的预测值则较为分散,相关性较差;SWR-SVM 方法所建模型对高降解率区域的预测能力则出现了严重低估,致使最终的建模质量不高。SWR-MLR 方法所建模型对训练集项目的估计能力较好,但对测试集项目的预测能力不足。总体而言,尽管 5 min 内二肽降解率定量预测模型能够提供一些参考信息,但模型质量有待改进,可能与它的降解机制相关。

表 1 不同时间段下 QSAR 模型的统计参数
Tab.1 Statistical parameter of QSAR models for different time.

时间/min	方法	<i>n</i>	<i>m</i>	模型			验证		测试	
				<i>R</i> ²	SD	<i>F</i>	<i>Q</i> ²	预测残差平方和	<i>n</i>	<i>R</i> ²
5	PLS	156	7	0.607	0.185	32.583	0.365	8.165	53	0.513
	SWR-SVM	156	7	0.395	0.229	8.602	0.354	8.314	53	0.259
	SVM	156	390	0.506			0.319	8.753	53	0.452
	SWR-RF	156	7	0.614	0.183	33.268	0.367	8.144	53	0.502
	RF	156	390	0.616			0.351	8.347	53	0.487
	SWR-MLR	156	7	0.515	0.205	22.405	0.455	7.004	53	0.406
30	PLS	156	8	0.703	0.189	43.574	0.508	8.698	53	0.514
	SWR-SVM	156	8	0.670	0.199	34.674	0.630	6.547	53	0.453
	SVM	156	390	0.661			0.437	9.955	53	0.358
	SWR-RF	156	8	0.707	0.188	44.264	0.498	8.870	53	0.537
	RF	156	390	0.707			0.471	9.347	53	0.530
	SWR-MLR	156	8	0.677	0.197	38.495	0.634	6.467	53	0.470
60	PLS	156	8	0.764	0.183	59.320	0.593	8.440	53	0.627
	SWR-SVM	156	8	0.682	0.212	41.803	0.657	7.103	53	0.593
	SVM	156	390	0.744			0.574	8.832	53	0.549
	SWR-RF	156	8	0.769	0.181	60.768	0.598	8.337	53	0.610
	RF	156	390	0.769			0.579	8.734	53	0.618
	SWR-MLR	156	8	0.689	0.210	40.682	0.644	7.384	53	0.582

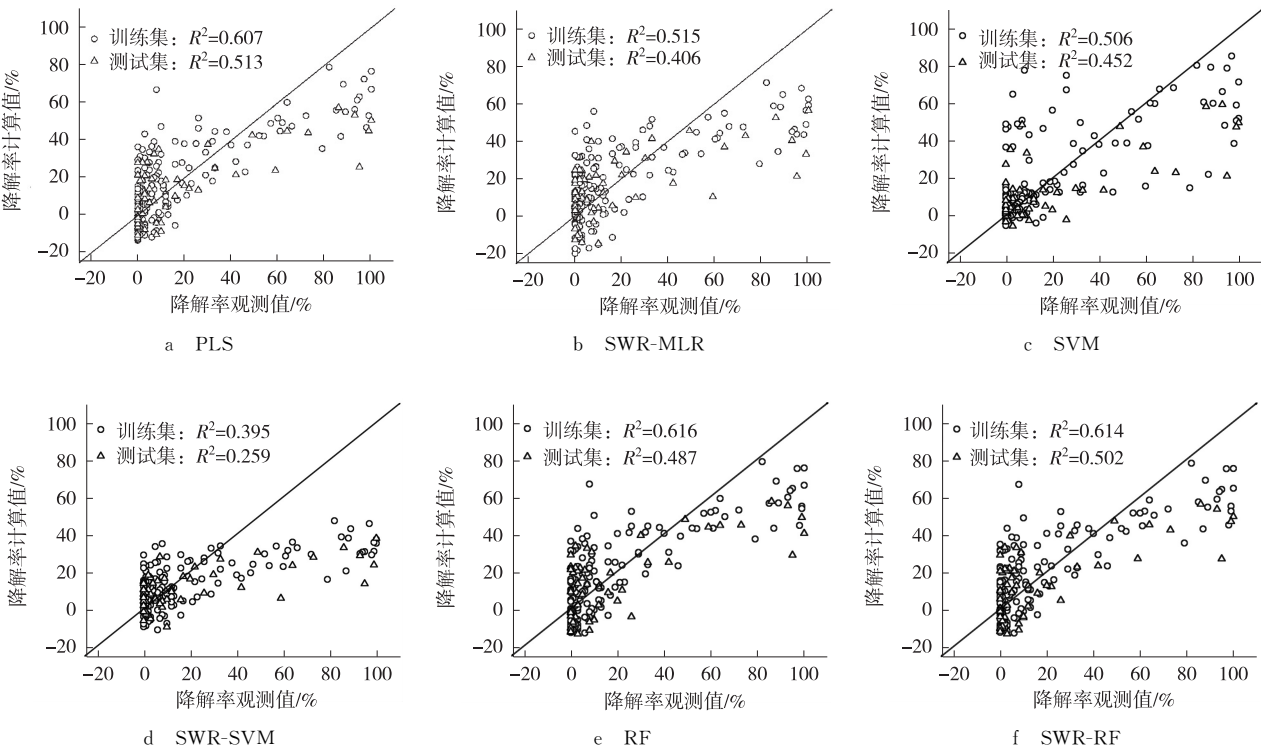


图 2 观测并计算 5 min 内不同 QSAR 模型的降解速率
Fig.2 Observed and calculated degradation rate from different QSAR models for 5 min

基于 SWR-MLR 所建定量预测模型的相关系数,计算了 20 种天然氨基酸对降解率的贡献。图 3a 显示二肽 1 位为精氨酸(Arg,R)、甲硫氨酸(Met,M)、赖氨酸(Lys,K)和酪氨酸(Tyr,Y),2 位为苯丙氨酸(Phe,F)、色氨酸(Trp,W)和酪氨酸时比较容易降解,例如二肽 YY, YF 在 5 min 的降解率高达 100.0%, FW 的降解率为 99.0%。相反,如果 1 位为天冬氨酸(Asp,D)、甘氨酸(Gly,G)、谷氨酸(Glu,E)和丙氨酸(Ala,A),2 位为天冬氨酸(Asp,D)、谷氨酸(Glu,E)和半胱氨酸(Cys,C)则不易发生降解,例如二肽 GD, GE 在 5 min 内的降解率为 0.0%。

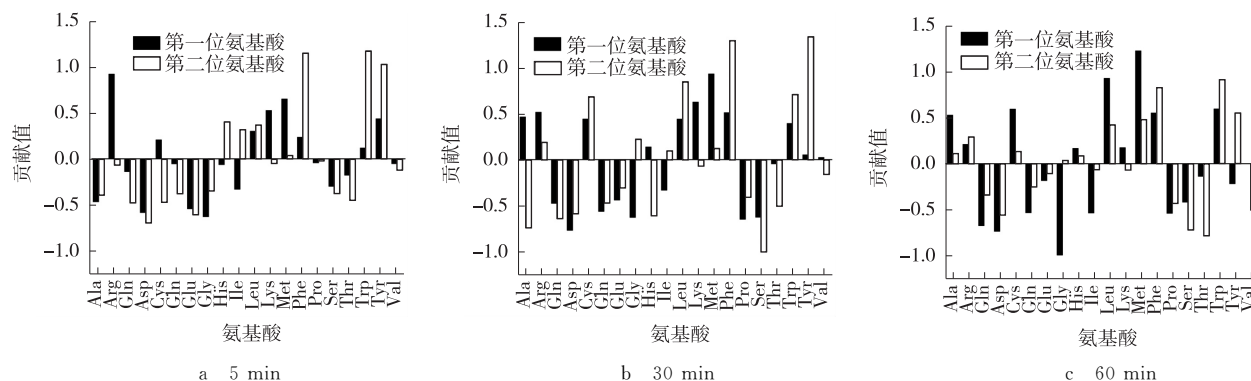


图 3 筛选变量的差异膨胀因子用于构建不同时间段的 QSAR 模型

Fig. 3 Variance inflation factors for screened variables used to construct QSAR model for different time

2.2 对 30 min 降解率的建模结果分析

对于二肽 30 min 降解率的数据集,逐步回归从初始变量中筛选出 8 个重要变量,图 1b 显示各有 4 个变量表征二肽的 1,2 位氨基酸,所有变量的 VIF 均小于 4,说明 SWR 所选变量具有良好的独立性。各种方法对该数据集的建模结果详见表 1,其中 PLS 方法模型由 8 个主成分构建。

对比 5 min 降解率数据集的建模结果,30 min 降解率数据集的定量预测模型质量有明显改善,所有模型均具有良好的相关性($R^2 > 0.66$, $Q^2 > 0.43$)与外部预测能力($R^2 > 0.35$)。特别是由 SWR-RF 和 RF 所建立的预测模型,训练集和测试集样本的实验观测值与计算值的相关性分别大于 0.70 和 0.53,具有良好的内部估计能力与外部预测能力。SWR-MLR 和 SWR-SVM 所建模型的稳定性较好($Q^2 > 0.63$),但外部预测能力(R^2 约为 0.46)还有待进一步改善。PLS 方法所建模型比较均衡,对内部建模样本($R^2 = 0.703$, $Q^2 = 0.508$)和外部测试样本($R^2 = 0.514$)预测值的相关性均较好。然而,尽管 SVM 方法对内部样本的估计能力较好,但稳定性($Q^2 = 0.437$)和外部预测能力($R^2 = 0.358$)均不如其他建模方法。

二肽 30 min 降解率实验观测值与计算值的相关性图(图 4)显示,SVM 方法所建模型的预测值离散度较大,特别是对外部测试集的预测能力不佳。其他 5 种方法所建模型的预测值与观测值的相关性较好,但是对外部测试集的预测能力还有待改善。

氨基酸贡献图(图 3b)显示甲硫氨酸、赖氨酸和苯丙氨酸在二肽的 1 位,以及苯丙氨酸、酪氨酸和亮氨酸(Leu,L)在二肽的 2 位容易发生降解,例如二肽 FY 的降解率为 100.0%。然而,二肽的 1 位为天冬氨酸、脯氨酸(Pro,P)和丝氨酸(Ser,S),2 位为丝氨酸、丙氨酸和谷氨酰胺(Gln,Q)则对提高稳定性有利,例如 PQ 的降解率为 0.0%。

2.3 对 60 min 降解率的建模结果分析

针对二肽 60 min 降解率的数据集,与 30 min 降解率的结果比较类似,逐步回归各筛选 4 个变量表征二肽的 1,2 位氨基酸(图 1c),所有变量的 VIF 均小于 4,表明所选变量所建预测模型能够有效避免过拟合问题。各种方法对该数据集的建模结果详见表 1。在 30 min 降解率数据集的建模基础上,各种方法对 60 min 降解率数据集所建模型的质量有进一步提高,所有模型对训练集样本的估计能力($R^2 > 0.68$)和外部测试集样本的预测能力($R^2 > 0.54$)均有改善,特别是预测模型的稳定性提升比较明显($Q^2 > 0.57$),表明预测模型能够很好用于二肽稳定性的研究。对比 6 种定量预测模型,PLS 所建预测模型不管是对内部训练集样本($R^2 = 0.764$, $Q^2 = 0.593$),还是外部测试集样本($R^2 = 0.627$),均具有良好的预测能力。SWR-RF 和 RF 所建模型的估计能力($R^2 = 0.769$)优

于PLS方法,对外部样本的预测能力也比较接近(R^2 为0.61左右)。由SVM和SWR-SVM所建模型的质量提升非常明显,特别是对外部样本的预测能力提高了10%以上。SWR-MLR所建模型的质量稍逊于其它几种方法,与最优模型相比,对内部样本的估计能力($R^2=0.689, Q^2=0.644$)相差不超过7%,对外部测试集的预测能力($R^2=0.582$)相差小于5%,表明该模型本身已具备良好的预测能力,而且基于其相关系数的氨基酸贡献可以为多肽稳定性研究提供直接信息。

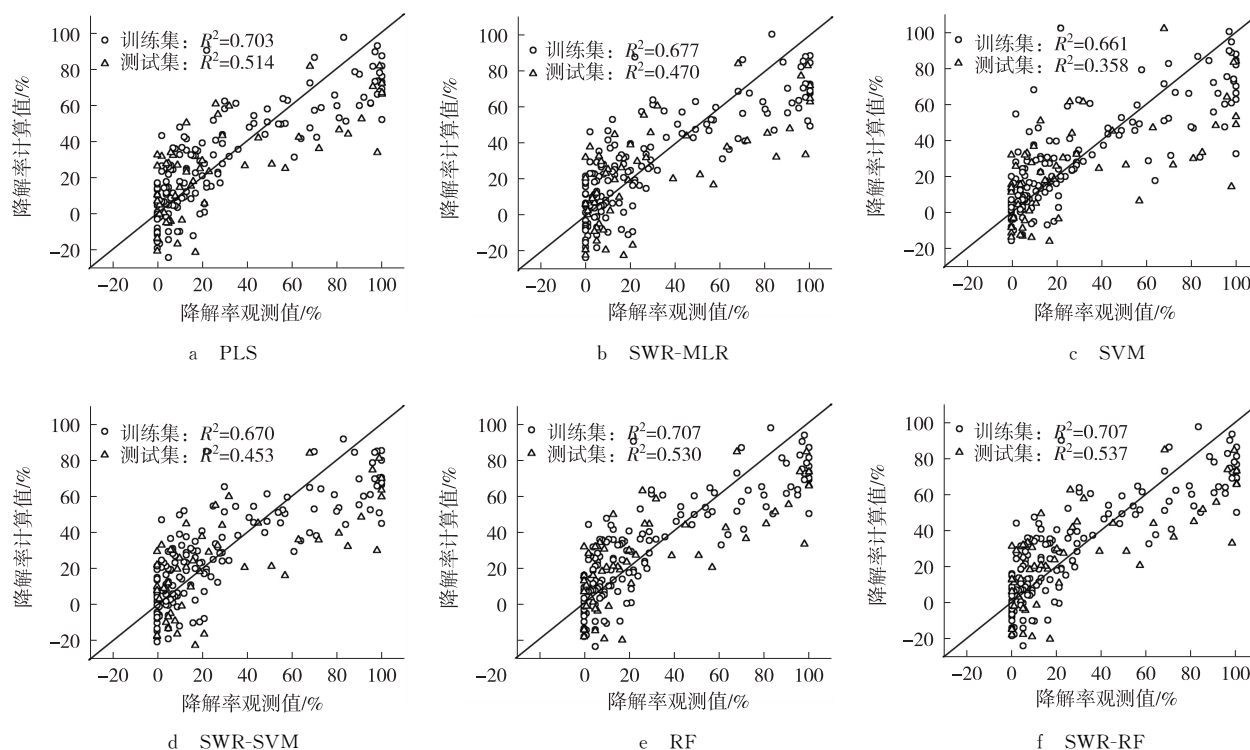


图4 观察并计算30 min内不同QSAR模型的降解率

Fig. 4 Observed and calculated degradation rate from different QSAR models for 30 min

二肽60 min降解率氨基酸贡献图与30 min降解率数据集基本相似,但贡献值稍有差异(图3c),如甲硫氨酸、亮氨酸和苯丙氨酸在二肽的1位,以及色氨酸、苯丙氨酸和酪氨酸在二肽的2位较容易发生降解,如FY的降解率为100.0%。相反地,二肽的1位为甘氨酸、天冬氨酸和谷氨酰胺,2位为苏氨酸(Thr, T)和天冬氨酸则能够提高稳定性,例如GD的降解率为0.0%。

本研究对二肽不同时间降解率所建立的定量预测模型显示,二肽的降解速率比并非线性,初始阶段的降解率与理化性质的线性关系较差,而60 min的降解率则与理化性质存在较好的线性相关性(图5)。此外,由于实验测定误差也影响所建定量预测模型的质量,例如二肽YY, YF在5, 30, 60 min的降解率分别为100%, 95%, 100%。整体而言,本文建立的定量预测模型,特别是对二肽在30, 60 min降解率所建的定量预测模型,要优于文献[10]的最优模型($R^2=0.578$),并且通过多种方法构建定量预测模型,确保了结果的准确性,能够知道高稳定性二肽分子的合理设计与虚拟筛选。

3 结论

本研究将SWR筛选变量与SVM, RF和MLR有机结合,构建209个二肽分子在5, 30, 60 min降解率的定量预测模型,结果显示SWR筛选变量独立性较好,能够有效避免预测模型的过拟合问题。本研究使用6种建模方法,对二肽60 min降解率数据集所建模型具有良好的估计能力($R^2>0.68, Q^2>0.57$)与预测能力($R^2>0.54$),明显优于文献[10]所建最优模型的预测能力($R^2=0.578$)。此外,基于MLR系数计算的氨基酸贡献能够发现影响二肽稳定性的重要氨基酸,为高稳定性二肽的设计提供直接线索。总之,本研究建立的预测模型方法简单,物理意义明确,对高稳定性二肽分子的设计与筛选具有理论指导作用。

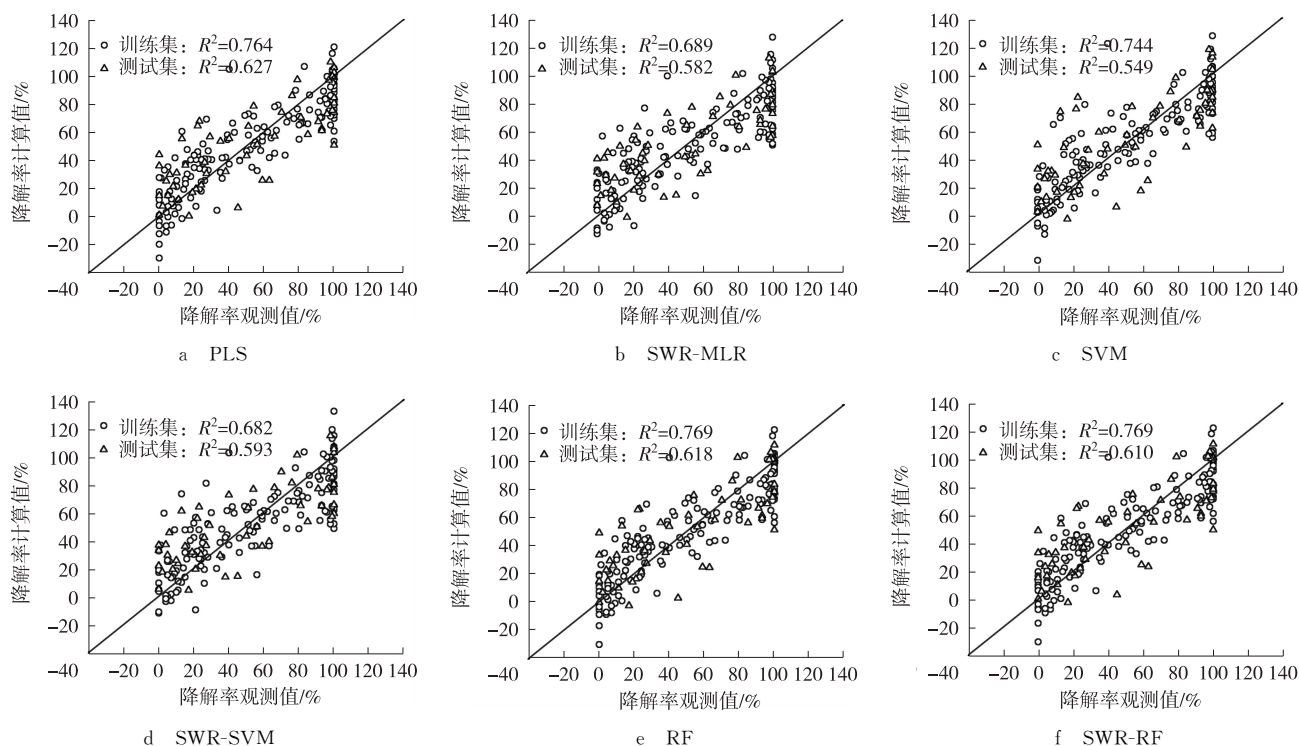


图 5 观测并计算 60 min 内不同 QSAR 模型的降解率

Fig. 5 Observed and calculated degradation rate from different QSAR models for 60 min

参考文献:

- [1] 余惠敏. 抗真菌药物的发展与研究现状[J]. 北方药学, 2012, 9(7): 26-27.
YU H M. Development and research status of antifungal drugs[J]. Northern Medicine, 2012, 9(7): 26-27.
- [2] 许亚平, 曹华. 多肽药物的研究及应用进展[J]. 广东药科大学学报, 2010, 26(6): 653-657.
XU Y P, CAO H. Progress in the research and application of peptide drugs[J]. Journal of Guangdong Pharmaceutical University, 2010, 26(6): 653-657.
- [3] ANAGNOSTIS P, ATHYROS V G, ADAMIDOU F, et al. Glucagon-like peptide-1-based therapies and cardiovascular disease: looking beyond glycaemic control. [J]. Diabetes Obesity & Metabolism, 2015, 13(4): 302-312.
- [4] 李玲, 洪战英, 董昕, 等. 多肽类药物的质量控制及药动学研究进展[J]. 药学服务与研究, 2013, 13(6): 405-409.
LI L, HONG Z Y, DONG X, et al. Progress in quality control and pharmacokinetics of polypeptide drugs[J]. Pharmaceutical Services and Research, 2013, 13(6): 405-409.
- [5] DENG P Y, LI Y J. Calcitonin gene-related peptide and hypertension. [J]. Peptides, 2005, 26(9): 1676-1685.
- [6] 姚金凤, 白露, 宋亚芳, 等. 多肽类药物代谢研究进展[J]. 中国药理学通报, 2013, 29(7): 895-899.
YAO J F, BAI L, SONG Y F, et al. Progress in the study of peptide drug metabolism[J]. Chinese Journal of Pharmacology, 2013, 29(7): 895-899.
- [7] 陈红丽, 王永学, 郭伟云, 等. 提高 PLGA 微球载蛋白多肽类药物稳定性添加剂的研究进展[J]. 国际生物医学工程杂志, 2012, 35(3): 185-188.
CHEN H L, WANG Y X, GUO W Y, et al. Research progress in improving stability additives of PLGA microsphere-loaded peptide polypeptides [J]. International Biomedical Engineering Journal, 2012, 35(3): 185-188.
- [8] TROPSHA A. Best practices for QSAR model development, validation and exploitation[J]. Molecular Informatics, 2010, 29(6/7): 476-488.
- [9] WU J, ALUKO R E. Quantitative structure-activity relationship study of bitter di- and tri-peptides including relationship with angiotensin I-converting enzyme inhibitory activity[J]. Journal of Peptide Science, 2007, 13(1): 63.
- [10] FOLTZ M, VAN B L, KLAFFKE W, et al. Modeling of the relationship between dipeptide structure and dipeptide stability, permeability, and ACE inhibitory activity [J]. Journal of Food Science, 2010, 74(7): H243-H251.
- [11] KAWASHIMA S, OGATA H, KANEHISA M. AAindex: amino acid index database[J]. Nucleic Acids Research, 1999, 27(1): 368-369.
- [12] MILLER A. The convergence of Efron's stepwise regression algorithm[J]. American Statistician, 1996, 50(2): 180-181.
- [13] CRANEY T A, SURLS J G. Model-dependent variance

- inflation factor cutoff values[J]. Quality Engineering, 2002, 14(3): 391-403.
- [14] EBERLY L E. Multiple linear regression[J]. Methods in Molecular Biology, 2007, 404(2): 165.
- [15] HEARST M A. Support vector machines[J]. IEEE Intelligent Systems and their Applications, 2002, 13(4): 18-28.
- [16] ARCHER K J, KIMES R V. Empirical characterization of random forest variable importance measures[J]. Computational Statistics & Data Analysis, 2008, 52(4): 2249-2260.
- [17] GELADI P, KOWALSKI B R. Partial least-squares regression; a tutorial[J]. Analytica Chimica Acta, 1985, 185(86): 1-17.

Studies on Quantitative Prediction Model of Dipeptide Stability Based on Amino Acid Properties

WANG Yuanqiang, GUO Haiqiong, WANG Yuxuan, ZHANG Yuping, ZHANG Ya, LIN Zhihua

(Chongqing Key Lab of Target Based Drug Screening and Activity Evaluation,

Chongqing Key lab of Medicinal chemistry and Molecular Pharmacology,

Chongqing University of Technology, Chongqing 400054, China)

Abstract: [Purposes] Dipeptide has the advantages of high biological activity and easy synthesis, but it has poor metabolic stability in the body and is easily degraded. Therefore, based on the properties of amino acids, the quantitative prediction of the stability of dipeptides was studied, which could provide a theoretical basis for the design of stable dipeptide molecules. [Methods] Uses partial least squares based on the degradation rate of 209 dipeptide molecules at different time periods (5, 30, 60 minutes), and combines stepwise regression screening variables with support vector machines, random forest, and multiple linear regression. A quantitative prediction model was established between the degradation rate of peptides and the physicochemical properties of amino acids. [Findings] The most notable model is the 60-minute degradation rate of dipeptides, with good estimation ability ($R^2 > 0.68$, $Q^2 > 0.57$) and predictability ($R^2 > 0.54$) for training sets and test sets, which can effectively predict The degradation rate of the peptide molecule; and the amino acid contribution calculated based on the multiple linear regression coefficients can find important amino acids that affect the stability of the dipeptide and can guide the rational design of the highly stable dipeptide molecule. [Conclusions] The prediction model established in this study is simple in method and clear in physical meaning. Various methods can obtain more ideal prediction models to ensure the accuracy of the prediction results. It can be used to guide the design and screening of dipeptide molecules with good stability.

Keywords: dipeptide; stability; degradation rate; quantitative prediction model; rational design

(责任编辑 许 甲)