国家自然基金(No. 30830090，60873103)资助
重庆理工大学优秀著作出版基金资助

蛋白质多肽序列表征方法及其应用

DANBAIZHI DUOTAI
XULIE BIAOZHENG FANGFA JIQI YINGYONG

舒 茂　林治华　杨 力 \ 著

西南交通大学出版社
Http://press.swjtu.edu.cn

蛋白质多肽序列表征方法及其应用

国家自然基金（No. 30830090，60873103）资助
重庆理工大学优秀著作出版基金资助

蛋白质多肽序列表征方法及其应用

舒　茂　林治华　杨　力　著

西南交通大学出版社
·成都·

内 容 简 介

定量构效关系是一种借助分子的理化性质参数或结构参数，以数学和统计学手段定量研究结构与性质的方法。通过计算建模，可以得到实验方法无法得到的资料，因此该方法广泛应用于药物、农药、化学毒剂等生物活性分子的合理设计。由于定量关系研究涉及面广，本书主要研究蛋白质多肽的定量构效关系。书中具体介绍了氨基酸结构表征方法、一般统计建模方法、多肽及蛋白质定量关系研究，图文并茂，实例生动。

本书适用于生物化学、药学及蛋白质组学相关专业的研究人员以及相关专业的学生使用。

图书在版编目（CIP）数据

蛋白质多肽序列表征方法及其应用 / 舒茂，林治华，杨力著. —成都：西南交通大学出版社，2010.10
ISBN 978-7-5643-0894-0

Ⅰ. ①蛋… Ⅱ. ①舒…②林… ③杨… Ⅲ. ①蛋白质－多肽－研究 Ⅳ. ①Q51

中国版本图书馆CIP数据核字（2010）第180852号

蛋白质多肽序列表征方法及其应用

舒 茂 林治华 杨 力 著

*

责任编辑 牛 君
特邀编辑 刘 恒
封面设计 何东琳设计工作室
西南交通大学出版社出版发行
成都二环路北一段111号 邮政编码：610031 发行部电话：028-87600564
http://press.swjtu.edu.cn
成都中铁二局永经堂印务有限责任公司印刷

*

成品尺寸：170 mm×230 mm 印张：11.25
字数：196千字
2010年10月第1版 2010年10月第1次印刷
ISBN 978-7-5643-0894-0
定价：33.00元

前 言

肽与蛋白质的结构表征是其定量构效关系（quantitative structure-activity relationship，QSAR）研究的前提和重要内容。由于肽和蛋白质的空间结构及功能信息隐藏于一级结构即氨基酸序列中，因此，氨基酸的结构信息对肽及蛋白质的定量构效关系研究至关重要。本书从氨基酸的结构特征出发，构建了两种全新的氨基酸结构表征体系，即 VHESH 和 VSTPV。VHESH（principal component score vector of hydrophobic，electronic，steric，and hydrogen bond properties）来源于 20 种天然氨基酸的 113 种物理化学性质，通过对其中 50 种疏水性质、23 种电性性质、35 种立体性质和 5 种氢键性质分别进行主成分特征提取而产生，其中 $VHSE_1$ 和 $VHSE_2$ 代表氨基酸疏水特性，$VHSE_3$～$VHSE_6$ 代表氨基酸电性特性，$VHSE_7$ 和 $VHSE_8$ 代表氨基酸的立体特性，$VHSE_9$ 和 $VHSE_{10}$ 代表氨基酸氢键供体和受体特性。VSTPV（principal component score vector of structural and topological variables）则来源于 166 种天然及非天然氨基酸的 85 种拓扑结构信息，并经主成分特征提取而产生。与 z-scale 等其他氨基酸描述子比较，VHESH 具有物理化学意义明确、表征能力强、结果易解释等优点；而基于氨基酸拓扑结构性质的 VSTPV 则具有计算方法简便、不依赖实验数据以及拓展性能好等优点。

在肽定量构效关系研究中，将 VHESH 和 VSTPV 用于血管紧张素转化酶抑制剂、后叶催产素、人类 1 型双载蛋白 SH3 结构域亲和肽、阳离子抗菌肽及细胞穿膜肽的定量构效关系研究，都取得了较好的构效关系建模结果。基于 VHESH 表征方法的构效关系研究发现：血管紧张素转化酶抑制剂第 2 残基电性与疏水性及第 1 残基立体等性质与生物活性呈正相关关系，而其第 1 残基的电性等性质则与活性呈负相关关系；后叶催产素第 1 残基电性及疏水性质和第 3 残基立体及氢键性质与其生物活性呈显著正相关关系，而第 2 残基疏水、电性及立体性质与其活性呈负相关关系；分析影响人类 1 型双载蛋白 SH3 结构域亲和肽亲和性关键作用力可知，第 P_{-3} 与第 P_2 之间残基（含 P_{-3} 与

P_2 残基）的相应性质对亲和活性影响较为显著；阳离子抗菌肽第 3 残基电性性质，第 6、7 和 12 残基立体性质以及第 11 和 12 残基的疏水性与抗菌效价呈正相关关系，而第 6、10 和 12 残基电性性质则与抗菌效价呈显著的负相关关系；细胞穿膜肽的相关残基的物化性质及拓扑性质对其穿膜性能影响较大。应用 VSTPV 表征方法对以上体系进行构效关系研究亦取得了较优的建模和预测结果，且得出影响活性关键氨基酸位点与 VHESH 模型结果基本吻合。在以上研究基础上，根据最优定量构效关系模型，在模型应用域范围内分别设计了一系列全新分子，其预测活性与各体系最高预测活性相比均有不同程度提高。

将 VSTPV 应用于含非天然氨基酸肽衍生物体系即血管舒缓激肽促进剂、牛乳清蛋白水解肽和弹性蛋白酶模拟底物的定量构效关系研究，取得了较好的结果。研究表明，血管舒缓激肽促进剂分子的第 2、3 残基相关拓扑信息与其生物活性呈强相关；牛乳铁蛋白水解肽的第 6、8 残基拓扑性质与其生物活性关系密切；弹性蛋白质模拟底物 A、B 残基部分变量的二次项和交互项对酶催化反应影响很大。

应用定量构效关系相关理论和方法对蛋白质特性及功能预测进行了研究。基于 VHESH 和 VSTPV 结构表征基础，对人免疫缺陷病毒蛋白酶裂解位点（HIV PR）、蛋白质磷酸化位点和蛋白质与 RNA 相互作用位点进行预测及特异性分析，取得了优于其他方法的预测结果。研究显示，HIV PR 的第 1、2、4、5 和 6 残基的立体、氢键、电性及疏水性质或对应的拓扑性质是 HIV PR 被识别重要因素；磷酸化位点序列的 P_{-3} 位点物化性质（VHESH）及其拓扑性质（VSTPV）对 S、T 和 Y 位点磷酸化影响最大；与 RNA 相互作用的蛋白质序列第 2、5、6 残基立体、疏性、电性和拓扑信息对 RNA 和蛋白质相互作用位点影响较大。

构效关系建模方法与技术是定量构效关系研究的一个重要内容。本书比较了多元线性回归（MLR）、偏最小二乘（PLS）、线性判别分析（LDA）及支持向量机（SVM）等方法在肽及蛋白质结构与功能关系研究中的应用。结果表明，MLR 在满足相关条件前提下，通常可以取得较好结果；PLS 可较好地解决变量数较多且存在多重共线性的情况；LDA 用于模式识别效果好，模型易解释；SVM 能较好地解决小样本、非线性、高维数和局部最小等问题。此外，为提高模型质量，采用多元线性逐步回归（SMR）、遗传算法（GA）筛选变量。研究发现，这两种方法能较好地删除原始变量中的噪音信息。

模型质量评价及其应用域现已成为建模方法学中的一个关键性问题。本

书将全部样本划分为训练集和预测集两个部分，由训练集样本建立 QSAR 模型，通过内部和外部双重验证来对模型进行质量评价。采用的内部验证方法有留一法（leave one out，LOO）、留组法（leave 1/*n* out，LNO）、留多法（leave many out，LMO）以及 Y 随机排序验证（Y random permutations test）。在内部验证基础上，通过多种评价函数对模型的外部预测能力进行评价，以确保模型的真实有效性。在此基础上，以样本的 X 空间标准化模型距离为依据确定了模型的应用域，避免模型外推后给活性预测带来的较大误差及不确定性。

蛋白质多肽的定量构效关系研究领域十分宽，发展十分迅速，各文献中有丰富的实例。由于篇幅的限制，书中不能一一介绍。

本书的编写得到了重庆理工大学优秀著作出版基金及国家自然基金（No.30830090，60873103）资助。

由于作者的水平有限，书中难免存在缺点和错误，敬请读者批评指正。

著 者

2010 年 7 月

目 录

1 绪 论

物质的许多物理化学性质以及生物学性质都是以分子为主体来表示和解释，分子结构一旦确定，其性质也随之而定。分子基本结构特征与其生理活性之间的关系，可以通过统计学方法对隐藏在大量实验结果中的信息和规律进行数据挖掘，并反过来指导实验过程。定量构效关系（quantitative structure-activity relationship，QSAR）就是一种借助分子的理化性质参数或结构参数，以数学和统计学手段定量研究有机小分子与生物大分子相互作用，有机小分子在生物体内吸收、分布、代谢、排泄等生理相关性质的方法。这种方法广泛应用于生物活性分子的合理设计。

蛋白质与多肽是生物体内非常重要的活性物质，由两个以上的氨基酸（amino acid，AA）分子通过酰胺键互相连接而成。蛋白质和肽不仅参与生物体的物质组成，调节生物体内各种生化反应以及跨膜运输等功能，同时还担任了诸如抗体、受体、激素等许多生物学功能。此外，多肽与免疫调节直接相关，是机体完成免疫功能和进行免疫调节的重要活性物质。蛋白质诸多生物学功能与特定肽链氨基酸排列顺序之间有着十分重要的关系，同时在生物体内直接发挥生物学功能的也有相当一部分属于肽类物质。因此研究多肽及蛋白质的定量构效关系对于研究其在生物体内的作用机理，以及肽类新药的研究与开发，同时对蛋白质结构与功能研究都具有十分重要的意义。鉴于此，近年来对多肽及蛋白质的结构与功能关系研究引起了国内外化学、药学、生物、医学甚至农学工作者的极大关注和浓厚兴趣。

1.1 定量构效关系研究概述

众所周知，相关学科的发展大大促进药物研究与开发，分子生物学（特

别是基因组学、蛋白质组学及代谢组学）、生物信息学与计算机辅助药物分子设计、组合化学、高通量筛选的快速发展已成为现代药学研究与开发的四大技术支柱，有可能缩短药物开发周期，提高药物开发效率，节约研究成本。药物分子设计涉及众多基础学科共同参与。早期的药物分子设计主要以经验设计为主，主要通过研究者自己的专业知识和直觉，根据一些实验现象来设计和优化药物分子，这不利于药物开发过程可控性。近年来，伴随着信息和计算科学发展而产生的计算机辅助药物分子设计学（CADD）成为了现代药物设计的主流，大大提高了新药设计效率和成功率[1-3]。CADD 主要由以下几方面组成：基于受体结构的定量构效关系、从头药物设计、虚拟高通量筛选和虚拟组合化学库设计。定量构效关系属于经典的 CADD 研究方法，在早期的药物设计中，定量构效关系方法占据主导地位。由于计算机计算能力提高和众多生物大分子三维结构准确测定，基于大分子结构药物设计逐渐取代了定量构效关系在药物设计领域的主导地位，但是定量构效关系在药学研究中仍然有一定研究地位。随着现代化学生物学、分子图形学及计算机科学等相关学科发展，定量构效关系也被注入新内涵。现在定量构效关系不仅仅是针对药物设计领域的 QSAR 研究，还包括定量结构-理化性质相关（QSPR）、定量结构-毒性相关（QSTR）、定量结构-色谱保留相关（QSRR）、定量结构-谱学性质相关（QSSR）等诸多方面。随着研究不断深入和扩展，如今 QSAR 研究已经成为化学、环境化学、计算化学、生物学、生物信息学、药学等多学科领域的一个前沿课题[4-6]。

1.1.1　定量构效关系研究进展

1. 二维定量构效关系研究

人们对化学结构有了初步认识后，就开始了化合物性质和分子结构之间关系的研究。QSAR 最早雏形是对有机化合物的结构与活性的研究，最早报道分子结构与特性关系的是 Cros 在 1863 年发现在哺乳动物体内醇类的毒性与水溶性呈反比关系[7]。直到上世纪 60 年代由 Hansch[8, 9, 11]和 Free[10]等人提出了自由能线性相关法（LFER）及亚结构分析法之后，QSAR 才正式走上药物设计的历史舞台。在他们开创性研究工作之后，相继涌现出许多新方法，QSAR 方法现已被广泛应用于各个研究领域，并日益成熟。

Hansch-Fujita 与 Free-Wilson 提出的模型是 2D QSAR 典型代表[9, 10, 12, 13]。Hansch-Fujita 模型以取代基对分子生物性质的影响是由其电性、立体性和疏水性三者中某些或全部因素变化引起，这三种效应往往同时起作用且彼此独立可加。Free-Wilson 模型以对于母体骨架相同的系列化合物，取代基在不同位置上的影响相互独立为假设。Hansch 方法和 Free-Wilson 方法的区别是：前者以物化参数解释这种贡献，虽然需要获取一些相关物理化学参数，有时候不易得到，但在一定范围内可以外延和内插设计新的分子；后者认为生物活性强弱是取代基本身影响，不需要物化参数，虽然不提供有关作用机理的信息，但可以预测化合物的活性，求出的系数直接反映基团对活性的贡献。Hansch-Fujita 与 Free-Wilson 方法已成功用于新药设计。

其他代表性的 2D QSAR 方法还有：分子连接性指数方法[14, 15]，该方法基于分子中原子之间连接性的二维结构特征编码，并通过代数运算得到的非经验参数，因此比较客观简便，不需通过实验方法求得，已成功地应用于合理组合库设计、虚拟筛选和药物设计等领域中，并发挥着十分重要的作用，由于没有明确的物化意义，应用受到一定的限制；电拓扑状态指数方法[16]，该指数是以分子图中每个原子图不变进行计算，将分子中成键原子的电子状态同它在整个分子骨架前后关系中的拓扑性质相关联，分子中有多少原子类型，相应就有多少种 E-状态指数，由这些指数共同描述整个分子，从而构成一组分子描述子；分子电性作用矢量等[17,18]，该矢量是基于分子内原子电性和分子二维拓扑结构的变量，业已取得较好研究结果。

2. 三维定量构效关系研究

经过十几年的发展，20 世纪 80 年代产生了能更加真实地反映药物与受体之间相互作用的三维定量构效关系研究方法。3D QSAR 模型是以药物分子的三维特征为基础，物化意义上更为明确，能直接反映药物分子作用过程中底物和受体之间的非共价键相互作用特征，主要处理药物分子三维空间中静电分布、氢键形成的能力和取向、立体性的配置和疏水性分布与生物活性之间的定量关系。现已产生众多 3D QSAR 模型化方法[1-5]，其中最具有代表性的是 Cramer 等[19]提出的比较分子场分析（Comparative molecular field analysis，CoMFA）方法，CoMFA 模型一方面映射出受体的拓扑结构和与药物结合的物化要求，同时可预测新化合物的活性。由于该方法是合适的校准策略分子对接问题的选择，现已得到广泛地应用与完善[20, 21]，但 CoMFA 方法受很多因素比如活性构象确定及分子叠合规则等的影响，结果不很稳定，同时具有一定

的应用限度[4]。为有效地避免在传统 CoMFA 中静电场和立体场函数形式引起的缺陷，Klebe 提出了 CoMFA 的扩展方法，即比较分子相似性指数分析（comparative molecular similarity indices analysis, CoMSIA）[22]，此外，CoMFA 正在不断地完善和发展[23, 24]。

3. 多维定量构效关系

Hopfinger 等[25]与 Albuquerque[26] 在其分子形状分析法基础上提出 4D QSAR 概念，以化合物分子各个构象、取向（ensemble profile）的集合为第四维，4D-QSAR 思想结合了构象、药效团和排布自由度来表达化合物生物活性。X 衍射分析已经证实，受体与配体结合时存在着一个诱导契合过程。4D-QSAR 中的活性构象并非最低能量构象，而是最优构象。不采取传统的构象系统搜索，而是在进行多温度分子热力学计算后，化合物构象进行集成采样，也就是对与化合物最低能量构象能量差上限范围内所有构象进行采样，构成构象集合作为化合物的最优能量构象集合。在这个过程中，由于化合物最优能量构象是以集合形式表现的，因此，4D QSAR 克服 3D QSAR 中存在的分子排列与构象选择等问题，已成功应用于 QSAR 研究[25-27]。

同时，为弥补3D和4D QSAR未考虑受体对配体的诱导锲合因素，2002年Vedani和Dober[28]提出Quasar方法[29]，开创了5D QSAR研究，第四维和前面一样，也是构象的集合，而第五维则是各种诱导契合的集合。5D QSAR较完全地考虑受体生物大分子结构，使药物设计更趋于合理。应用5D QSAR 可以提高配体排列程度，选择出一个合适的诱导契合模型。这种模拟过程可以识别一个简单的活性组配体和一个简单的诱导契合模型，但是却不适合分析对整个化合物贡献很小的实体。更重要的是，在5D QSAR 中应用的遗传算法不一定仅仅局限在给定的立体场、静电场、氢键场、分子亲脂势能、能量最低和线型等适应类型中，而且可以建立任何线形和各向异性模型。与4D QSAR 相比，5D QSAR 模型的优点不仅适合于小分子的活性研究，甚至当该化合物比配体分子大得多情况下都可以用给定的模型研究新化合物QSAR。另外，诱导契合中各种假设相互交互作用可以通过整个模拟过程实现。Vedani等[30]在前期研究基础上又提出6D QSAR概念，其第六维是指溶剂化效应。他对106个结构各异的雌激素受体配基的研究取得了优良的QSAR模型结果，发现雌激素受体的结合部位应具有大面积的疏水区，配体在结合时会有明显的去水合作用，因此，增加溶剂化效应能改善计算结果。6D QSAR 对其他药物定量构效关系研究是否取得有效结果，还有待进一步证明。

4. 二维与多维定量构效关系之间关系

从定量构效关系的提出到 6D-QSAR 的诞生，代表了今后 QSAR 研究的主要发展方向，定量构效关系目前仍处于不断发展、完善和方法学探索阶段，一些问题还需解决。到目前为止，2D-QSAR 仍然发挥着多维 QSAR 不可替代的重要作用，如药物在体内的吸收、分布、代谢、排泄、跨膜转运、生物利用度的 QSAR 研究，柔性较大化合物的 QSAR 研究，以及整体动物体内的生物活性数据测定利用等问题。因此，基于 2D QSAR 研究基础上的分子设计至今仍然具有强大的生命力和实际应用价值。在很多定量构效关系研究体系中，2D-QSAR 常常会取得优于多维 QSAR 的研究结果。除此之外，2D-QSAR 方法还具有计算相对简便，计算速度快，可以大批量处理化合物数据，而多维 QSAR 方法目前还难以实现，同时多维 QSAR 有时还需要 2D-QSAR 的结果作为依据。因此，多维 QSAR 方法出现和发展并不会替代原有 2D-QSAR 方法，它们相辅相成，互为补充。

5. 定量构效关系研究的共同特征

至今为止，已有多种方法用于定量构效关系研究，它们具有以下共同特征[31]：①假定化合物分子结构与生物活性或物化性质间存在一定相关性。②经分子结构表征，化合物分子结构可用适当的描述子表示。这些分子结构参数包括[32, 33]：脂水分配系数、疏水参数及HPLC保留值等；立体参数，如Taft立体效应参数及STERIMOL立体参数等；电性参数，如Hammett电性常数及Taft诱导效应参数等；其他混杂描述子，如实验特性、氢键线性溶解能、计算特性、拓扑参数、量子化学参数、光谱性质参数、指纹图谱描述子及指示变量等。③根据已知化合物结构-活性数据建立的函数，可内插或外推至新的化合物。目前定量构效关系研究核心仍然是方法学问题，包括分子结构表征、理论模型推导方法、函数关系建立和生物活性预测等[34,35]。

1.1.2 定量构效关系研究前景

随着计算机技术和分子生物学、分子药理学的快速发展，定量构效关系已从经典的二维定量构效关系发展到多维定量构效关系，QSAR 技术对有机合

成化学、药物化学及药物设计的发展起了巨大推动作用，使人们对药物配体-受体的结合过程有了更深入认识，这对于药物分子设计和先导化合物改造有十分重要的意义。随着化学、药物化学、生物化学及计算机科学等领域交叉作用及各领域专家通力合作，QSAR 方法将不断发展和完善，成为研究物质理化性质与生物活性以寻求分子解释的一个强有力的工具[36]。随着 QSAR 研究不断发展和不断丰富，不断地向环境科学、生命科学和计算机科学等其他学科渗透，定量构效关系研究将会成为多学科交叉领域的一个重要学术前沿。

1.2 肽定量构效关系研究进展

肽和蛋白质在生物体内具有非常重要的作用，它们是生物体内物质组成的基础，参与了生物体内各种生化反应和跨膜运输过程，是抗体、受体、激素等发挥生物学功能的前提，因此，对肽和蛋白质结构功能的研究引起了国内外化学、生物、药学、医学等方面专家的极大关注和浓厚兴趣。蛋白质的诸多生物学功能与特定肽链的氨基酸排列顺序亦有着十分密切的关系，因此研究肽的定量构效关系对于研究其在生物体内的作用机理，以及肽类新药的研究与开发，同时对蛋白质的结构与功能研究都具有十分重要的意义。多肽具有高活性、高选择性及副作用小等特点，现已成为药物研究热点之一。对肽类药物及其先导化合物的研究开发发现，仍然耗资巨大且效率低。因此，迫切需发展新的理论方法指导肽类药物开发。QSAR 方法已广泛应用于肽物理化学性质测定、肽营养保健品以及肽生物活性建模分析[37-39]。近年来，计算机辅助分子设计，在肽类化合物的研发中得到广泛应用。以计算机分子图形学、分子动力学和量子化学等进行构象分析，寻找多肽及类似物的药效团，进行 QSAR 研究，用各种方法设计有较高活性肽及拟肽，已成为国际上十分活跃的研究领域。

肽 QSAR 研究主要有以下 5 个步骤：①选择和设计一系列肽类似物；②合成肽类似物并进行生物活性测定；③肽类似物结构定量表征；④建立数学模型，确定化学结构与生物活性之间的函数关系；⑤新类似物活性预测以及新高活性类似物的设计。由于肽类物质结构相对复杂且柔性高，其功能特性涉及到序列中氨基酸位置、构成及其物理化学性质等因素，因此，多肽 QSAR 研究主要集中在多肽结构表征研究。氨基酸结构表征参数定量描述多肽化学

结构，是多肽 QSAR 研究中一个核心问题。

氨基酸结构表征方法最早由Sneath[40]用物化半定量参数得到。Kidera等[41]将统计学方法用于氨基酸性质变量预处理，Kidera收集到20种天然氨基酸188种性质并用因子分析法得到10个正交因子，并在预测或确定蛋白质和多肽高级结构研究中取得较好的结果。Hellberg等[42,43]在Kidera结构参数化研究基础上，使用主成分分析（principal component analysis, PCA）技术对天然氨基酸29种理化性质进行了信息压缩和提取，确立了PCA作为氨基酸描述子处理的基本工具。使用PCA所产生的3个显著主成分分别代表氨基酸疏水、立体和电性性质，并分别用z1、z2和z3来表示，并通过偏最小二乘方法（partial least square regression, PLS）成功地建立了一些肽QSAR模型，并先后在对催产素、胃酶抑素、血管舒缓激肽促进剂、苦味二肽等多个肽体系的QSAR研究中取得了较优结果。此后，Sandberg等[44]又将z-scales扩展到包括天然氨基酸在内的87种氨基酸，并进行拟肽QSAR研究。z标度被后人成功应用于肽活性预测[45]、蛋白质设计[46]、肽-蛋白质亲和性分析[47]、蛋白质稳定性判别[48]等多方面研究。

Collantes等[49]以肽链上氨基酸侧链全向表面积（isotropic surface area, ISA）和侧链所有原子净电荷指数（electronic charge index, ECI）为基础建立多肽3D QSAR模型，在血管紧张素转化酶抑制剂、苦味二肽、催产素、血管舒缓激肽促进剂及速激7肽的QSAR研究中取得较优结果，进而在组合肽库设计[50]、HLA-A*0201限制性CTL表位识别[51]、肽骨架电荷分布模拟[52]等方面研究取得了较好成效。但是由于ECI/ISA参数仅为二类性质变量，其包含信息量有限，因而难以有效描述复杂多肽的序列和结构特征。

Zaliani等[53]从氨基酸36个参数中提取出分子立体和电性等三维特征参数MS-WHIM标度，并将其用于苦味二肽和血管收缩素转化酶抑制剂等体系进行了定量构效关系研究，结果较好。采用MS-WHIM结构描述符建立的QSAR 模型物理意义不甚明确，但所建模型有很好的预测能力，且其可进一步表征非天然氨基酸，以此来解决含非天然氨基酸多肽定量描述问题。

除了z标度、ECI/ ISA指数和MS-WHIM标度以外，人们还从不同的角度定义了氨基酸描述子，如理论计算及量子化学参数性质[54-59]描述子，氨基酸侧链物理化学性质及拓扑结构性质[60-66]的描述子在肽序列模型与活性预测去得到较好的结果。但在天然氨基酸表征方面还是缺乏很好地反映氨基酸疏水、电性、立体及氢键等物化性质描述子，这需用能综合体现这些性质的表征方法。在非天氨基酸表征方面，由于其结构非常广泛，不易用物化性质来表征，如果用量化计算方法来表征会非常耗时，因此需要简单而快速的表征方法。

当对含有不同氨基酸数目的肽序列进行 QSAR 研究时，用上述的 z 标度、ISA/ECI 指数等描述子来对样本序列进行结构表征，势必会产生出不同数目的描述子变量，进而无法进行 QSAR 建模。Andersson 等[42]采用自交叉协方差（auto cross covariance, ACC）数据预处理技术，使每一个样本均产生出相同数目变量，从而使常规建模方法得以应用。该方法最初是用于 DNA、蛋白质和肽模式识别领域的研究，但很少应用于 QSAR 建模。同时，Andersson 等[42]将正交信号校正（orthogonal signal correction, OSC）技术应用到 QSAR 建模研究中，也取得了较好的结果。OSC 是探索性的应用于 QSAR 建模分析，其有效性必须借助于外部验证的结果来进行判断。

肽 QSAR 研究目前还没有较为成熟的方法学指导，各种方法和技术都还处于不断地尝试和发展之中。目前，肽 QSAR 研究需解决以下两方面问题 [67]：一是多肽结构表征，虽然现已出现各种定量分子结构的实验测定参数和理论计算参数方法，但因多肽类似物分子较大、结构复杂多样且柔性较高，如何定量表征多肽序列仍是肽 QSAR 研究中的难题之一；二是统计建模方法，多肽的 QSAR 建模过程中，普遍存在化合物个数少、结构描述参数多的问题，因此，需要完善现有统计建模方法。

伴随着生物信息学、分子模拟、分子图形学等学科的兴起，它们也给 QSAR 注入了新的活力，使其正朝向过程可视化、信息多元化、模型容易解释及计算性能高精度化方向发展。多肽 QSAR 会更多地用于指导先导化合物研究和开发，多肽类药物分子设计方法会更加完善和丰富。

1.3　蛋白质特性及功能预测研究进展

蛋白质结构与功能的研究是分子生物学核心内容之一，研究蛋白质结构与功能不仅具有重要理论意义，而且对生物技术的发展也具有重要实践意义[68, 69]。蛋白质功能由其结构决定，因此，要研究蛋白质功能则需要研究蛋白质结构。蛋白质结构预测可分为：序列分析，二级结构预测，三级结构预测。在结构预测研究中，一是根据经验势能参数的自由能最小化 [70, 71]，即假定蛋白质二级结构主要由邻近氨基酸残基的相互作用决定，通过分子动力学或其他方法计算，找出自由能最低构型。另一类是统计方法[72, 73]，即对已知结构的蛋白质进行统计分析，建立序列到结构映射模型，进而根据映射模型对未知