

文章编号:1000-582X(2003)08-0035-04

活性寡肽分子结构表达及生物功能预测

仇亮加¹,李波¹,李劲为¹,王小燕^{1,2},李志良^{1,2}

(1.重庆大学化学化工学院,重庆 400044; 2.重庆大学教育部重庆市生物医学工程重点实验室,重庆 400044)

摘要:基于点电荷作用原理,仅从寡肽一级结构出发,借助分子中各原子电负性及原子间距离,构建了能描述多肽分子结构的描述子参数——分子电负性距离矢量简称分子电距矢量。据此对24个速激七肽序列进行了定量结构活性相关(QSAR)研究,取得了与文献接近或更好的效果,同时从多元线性回归分析结果中获取了一些关于寡肽空间结构的信息,提出的电距矢量的计算比文献更为简便、更易获得,且活性相关性好的特点,可望在生物大分子的结构表征及活性预测方面有所作用。

关键词:定量结构活性相关;分子电边矢量;速激七肽;活性寡肽;结构表达;功能预测

中图分类号:R95; R969.1

文献标识码:A

人类基因组计划^[1]已于20世纪80年代中叶启动,取得了举世瞩目的巨大成就^[1]。随着现代生命科学前所未有的进展,使生命科学正从描述性向推理性、从宏观状态的研究到微观结构理论发展的趋势,特别是近年来有关蛋白质结构与功能关系的研究受到了国内外生物学家、化学家以及相关学科的研究者们足够的重视。国内王志新院士和来鲁华教授^[2]在这方面作了大量的研究,取得了较大的进展,特别是在蛋白质空间结构的预测方面作了不少工作。实际上,自从 Anfinsen 等^[3-4]提出蛋白质的一级结构完全决定了其三维空间结构的著名论断以来,根据蛋白质的氨基酸序列从理论上预测其相应的空间结构就成为分子生物学中一个重要的、迄今为止仍未解决的难题。1993年 Martin 等^[5]进一步阐明了蛋白质高级结构的信息仍全部包含在其一级结构中。近年来不少学者在这方面作了探讨^[2-7]。Raychaudhury 等^[7]仅从肽链的一维结构出发构建了描述氨基酸侧链拓扑形状与尺寸的结构参数并用于定量结构活性相关性研究得到了较满意的结果,但计算相对较繁琐。文章从肽链一级结构出发基于肽链中原子间的距离及各非氢原子的电负性构建了能描述肽链结构的分子电负性距离矢量(MEDV),据此,借助多元线性回归分析在分别成功建立二肽及五肽等寡肽的 QSAR 模型之后,又对24个速激七肽序列

进行了定量结构活性相关研究,取得了与文献接近或更好的效果。

1 原理与方法

“结构决定性质,性质是结构的反映”,这是化学中的普遍规律。化合物中各组成原子好比物理学中的带电荷的质点,其相互作用可按点电荷间的相互作用规律即库仑定律描述如下式(1):

$$E = \sum_{i=1}^n (q_i * q_j) / d_{ij} \propto (q_i * q_j) / d_{ij} \quad (1)$$

其中 q_i, q_j 分别表示第 i 、第 j 类原子的相对于碳原子的电负性。 d_{lm} 表示属于第 i 类和第 j 类的第 l 、 m 两原子间的相对距离,这里用原子间的键联距离相对与碳—碳单键的相对距离,其中相邻两原子的键联距离为两原子按其原子半键长的加和值,由于形成肽链的实际键长可由文献得到,因此肽链上部分是取得实际键长。称之为分子电负性距离矢量(简称分子电距矢量 MEDV 或 VMEE,记为 v)。

$$u_k = u_{ij} = \sum_{i=1}^n (q_i * q_j) / d_{ij} \propto (q_i * q_j) / d_{ij} \quad (1)$$

其中 v_k 表示第 k 个描述子变量。可计算出速激七肽 DwFwGLM(No.1)的 VMEE 矢量的10个变量,类似地,分别为 $v_1 = 0.48878$; $v_2 = 6.84849$; $v_3 = 5.3474$; $v_4 = 4.58012$; $v_5 = 26.081$; $v_6 = 45.6486$; $v_7 = 36.5308$; $v_8 = 29.0928$; $v_9 = 13.1319$; $v_{10} = 1.19976$ 。从而 DwFwGLM 的 VMEE 矢量为: $v = (0.48878, 6.84849,$

* 收稿日期:2003-04-01

基金项目:重庆市应用基础项目;国家春晖计划教育部启动基金;霍英东基金及国家新药基金资助课题。

作者简介:李波(1983-),男,广西桂林人,重庆大学本科生。

5.347 4, 4.580 12, 26.081, 45.648 6, 36.530 8, 29.092 8, 13.131 9, 1.199 76)'. 其它速激七肽的 ν 矢量计算方法与此相同。

2 结果与讨论

2.1 血管紧张素转化酶抑制剂的 QSAR 研究

为检验分子电边矢量 ν 对肽链结构的表征能力, 运用多元线性回归方法建立 58 个血管紧张素转化酶抑制剂活性与对应的分子电边矢量 ν 的 QSAR 模型,

表 1 QSAR 的统计特征

	bi	t-test	F-test	SSPR	PCC	bi	t-test	F-test	SSPR	PCC	bi	t-test	F-test	SSPR	PCC
01	1.389	1.062 8	1.129 6	1.937 1	0.600 8	1.125 0	0.752 6	0.566 5	1.270 5	0.469 8	1.135 2	0.776 5	0.603 0	1.293 6	0.481 3
02	-2.628	-0.616 3	0.379 8	0.651 3	-0.399 5	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
03	4.355	0.530 4	0.281 3	0.482 5	0.351 2	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
04	-7.970	-0.920 5	0.847 3	1.453 0	-0.545 5	-3.901 2	-0.394 0	0.155 2	0.348 2	-0.268 4	-4.226 9	-0.436 5	0.190 5	0.408 7	-0.294 9
05	-8.864	-0.652 3	0.425 5	0.729 7	-0.418 8	-5.546 9	-0.356 9	0.127 4	0.285 7	-0.244 7	-4.969 3	-0.326 9	0.106 9	0.229 3	-0.225 2
06	7.162	0.435 1	0.189 4	0.324 7	0.294 1	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
07	-0.929	-0.050 8	0.002 6	0.004 4	-0.035 9	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
08	10.635	0.401 3	0.161 0	0.276 1	0.273 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0
09	-23.427	-0.561 8	0.315 6	0.541 3	-0.369 2	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	-4.042 5	-0.086 7	0.007 5	0.016 1	-0.061 2
10	70.969	0.949 5	0.901 5	1.545 9	0.557 4	27.727 7	0.324 4	0.105 2	0.236 0	0.223 6	33.993 0	0.406 6	0.165 3	0.354 7	0.276 3
11	-81.551	-1.191 4	1.419 5	2.434 3	-0.644 3	-42.353 9	-0.541 1	0.292 7	0.656 6	-0.367 3	-44.903 0	-0.586 5	0.344 0	0.738 0	-0.383 1

说明: t 检验值, F 检验值, SSPR 偏回归平方和, PCC 偏相关系数, $R^2 = 0.8804$, $R = 0.9383$, $SD = 1.3095$, $U = 25.2369$, $Q = 3.4297$, t 检验的显著性水平 $P = 0.05$, $t = 4.303$

表 2 速激七肽序列及其 RPA 实测与计算活性

nCompd.	yBA	Cal.	Err.	Err%	Cal.	Err.	Err%	Cal.	Err.	Err	Cal.	Err.	Err
1DwFwGLM	5.23	4.40	0.83	15.87	4.748	-0.482	-9.22	4.782	-0.448	-8.57	4.746	-0.484	-9.25
2DYFwGLM	4	4.92	-0.92	-230	4.467	0.467	11.67	4.431	0.431	10.77	4.463	0.463	11.56
3DYwVwwR	7.65	6.10	1.55	20.26	6.937	-0.713	-9.32	7.010	-0.640	-8.37	6.952	-0.698	-9.12
4DYVwVwwR	4	4.25	-0.25	-6.25	4.373	0.373	9.33	4.366	0.366	9.15	4.341	0.341	8.52
5DYPVwwR	4.89	4.94	-0.05	-1.02	4.631	-0.259	-5.30	4.609	-0.281	-5.74	4.651	-0.239	-4.88
6DyvwVwwr	6.39	6.53	-0.14	-2.19	6.482	0.092	1.44	6.420	0.030	0.47	6.494	0.104	1.63
7DYwVwGwR	4.71	4.50	0.21	4.46	4.824	0.114	2.42	4.795	0.085	1.81	4.812	0.102	2.17
8DyvwVwgr	6.58	6.02	0.56	8.51	6.504	-0.076	-1.16	6.568	-0.012	-0.19	6.497	-0.083	-1.26
9DYwVwLR	6.72	6.28	0.44	6.55	6.834	0.114	1.70	6.805	0.085	1.27	6.822	0.102	1.52
10NYwVwwR	7.19	7.25	-0.06	-0.83	7.144	-0.046	-0.63	7.150	-0.040	-0.55	7.149	-0.041	-0.56
11DYVYYR	4	3.82	0.18	4.50	3.902	-0.098	-2.46	3.933	-0.067	-1.69	3.919	-0.081	-2.03

说明: "/" 为未给出; "*" 为预测活性

2.2 苦味二肽的 QSAR 研究

从文献[8]选取 48 个苦味二肽进一步来检验电边矢量 ν 对肽链的表征能力。首先选取电边矢量 ν 中的 10 个元素运用多元线性回归技术建立 48 个二肽电边矢量 ν 与其相应苦味阈值间的定量结构-活性相关模型(QSAR), 根据 QSAR 模型的标准回归系数可以看出电边矢量 ν 中第 1、2、6 三个元素对活性的贡献较大, 又特别是第 1 个元素的贡献最大, 而其余几个元素的贡献相对较小。而电边矢量 ν 的第 1 个元素表示的是所有相邻原子间的相互作用, 说明 48 个苦味二肽的生物活性阈值的大小主要取决于肽链的一级结构。于是作者只选取第 1、2、6 三个元素建立 QSAR 模

型, 其结果与选用电边矢量 ν 的十个元素建立 QSAR 模型的结果很相近。QSAR 模型的统计特征值列入表 1。表 2 列出了相应的观察值与计算值。本文用电边矢量 ν 计算的结果与文献[3-6]结果比较接近。

2.3 血管舒缓激肽增效剂(五肽)的 QSAR 研究

从文献中选取 31 个五肽(血管舒缓激肽增效剂)作为样本集建立 QSAR 模型。首先选取分子电边矢量 ν 中的 10 个元素运用多元线性回归技术建立 31 个五肽分子电边矢量 ν 与其相应活性间的 QSAR, 根据 QSAR 模型的标准回归系数可以看出分子电边矢量 ν 中第 1、5、6、7、8、9、10 七个元素对活性的贡献较大, 有趣的是而这几个元素中第 1 个元素的贡献没有第 5、

6、7、8、9、10几个元素的贡献大,这与二肽的结果是不同的。而分子电边矢量 ν 的第5、6、7、8、9、10几个元素表示的是分别相隔5、6、7、8、9、10个键的所有原子间的相互作用,为什么会出现相隔5、6、7、8、9、10个键的原子间相互作用的影响比相隔1、2、3、4个键的原子间相互作用的影响还大的这种现象呢?这可能是因为在形成高级结构以后,分别相隔5、6、7、8、9、10个键的所有原子间的实际距离由于形成维系高级结构的氢键或其它作用而拉近了,因而其间相互作用的影响就变得较大。这个结果从另一方面说明了:对于多肽,其生物活性不仅与其一级结构有关,而且与其高级结构有关,一级结构是基础,高级结构是影响活性的主要因素。于是作者选取第1、5、6、7、8、9、10七个元素建立QSAR模型,其结果与选用分子电边矢量 ν 的十个元素建立QSAR模型的结果很相近。表4列出了相应的观察值与计算值, QSAR模型的统计特征值列入表1。结果表明,生物活性的计算值与观察值吻合良好。文献[7]分别从所述31个五肽中选取表4前15个五肽作训练集运用偏最小二乘方法(PLS)建立QSAR模型,并对余下16个五肽的生物活性进行了预测,取得了较好的结果。为便于比较,作者亦从文献的31个五肽中选取同样的15个五肽作为样本集,直接运用多元线性回归(MLR)方法,籍助本文提出的分子电边矢量 ν 建立QSAR模型。根据上述对总体建模的标准回归系数所选出的七个元素建立QSAR模型, QSAR模型的统计特征值列入表1,表2列出了相应的观察值与计算值。本文用分子电边矢量 ν 计算的结果与文献[7]结果相比较接近或更好,但本文分子电边矢量 ν 的计算简便得多,只需知道肽链的一级结构便可计算出分子电边矢量 ν ,不需其它任何理化参数或理论参数。作者运用样本所建的QSAR模型对余下16个五肽的生物活性进行了预测,亦得到了较满意的结果。

2.4 速激七肽的QSAR研究

通常对肽链结构表征及QSAR研究时,大多是以氨基酸侧链为基础,用多肽中每个侧链氨基酸的电性、几何或立体参数来描述肽链的结构,取得了一些令人满意的进展。但这类方法有一个较大的弱点,就是随着肽链的增长、氨基酸侧链的增加,肽链结构描述参数将成倍地增加。为检验分子电边矢量 ν 对多肽链结构表征和生物活性预测能力,从文献中选取24个速激七肽作为样本集建立QSAR模型。首先选取分子电边矢量 ν 中的15个元素及指示变量六元素运用多元线性回归技术建立24个速激七肽分子电边矢量 ν 与其相应活性间的QSAR模型, QSAR的统计特征值列入表1。然后,运用分子电边矢量(MEEV)对所有24个序列作训练集采用逐步多元回归方法(SMR)建立定量结构活性相关(QSAR)模型,根据标准回归系数从分子电边矢量中选取14个(或15个)元素作描述变量建立QSAR模型,结果良好,相应统计特征值列入

表1。

用分子电边矢量 ν 计算的结果与文献[8]结果相比较接近或更好。但文章中分子电边矢量 ν 的计算较为简便,只需知道脱氧核糖核酸启动子序列链的一级结构便可计算出分子电边矢量 ν ,不需其它任何理化参数或理论参数。作者运用所建的QSAR模型对24个速激七肽序列的生物活性进行了预测,得到了较满意的结果。根据QSAR模型,从标准回归系数可以看出分子电边矢量 ν 中第1-4、5、6、8、9、10-15等元素对活性的贡献较大,有趣的是而这几个元素中第1个元素的贡献没有第2-5、6、8、9、10-15等几个元素的贡献大,这与二肽的结果是不同的。而分子电边矢量 ν 的第5、6、8、9、10-15几个元素表示的是分别相隔5、6、8、9、10-15个键的所有原子间的相互作用,为什么会出现相隔5、6、8、9、10-15个键的原子间相互作用的影响比相隔1、2、3、4、7个键的原子间相互作用的影响还大的这种现象呢?这可能是因为在形成高级结构以后,分别相隔5、6、8、9、10-15个键的所有原子间的实际距离由于形成维系高级结构的氢键或其它作用而拉近了,因而其间相互作用的影响就变得较大。这个结果从另一方面说明了:对于多肽,其生物活性不仅与其一级结构有关,而且与其高级结构有关,一级结构是基础,高级结构是影响活性的主要因素。于是作者选取第1、5、6、7、8、9、10七个元素建立QSAR模型,其结果与选用分子电边矢量 ν 的10个元素建立QSAR模型的结果很相近。QSAR模型的统计特征值列入表1,表2列出了相应的观察值与计算值。结果表明,生物活性的计算值与观察值吻合良好。笔者用分子电边矢量 ν 计算的结果与文献[4]结果相比较接近或更好,但文章中分子电边矢量 ν 的计算简便得多,只需知道肽链的一级结构便可计算出分子电边矢量 ν ,不需其它任何理化参数或理论参数。作者运用所建的方法对其它多肽的生物活性进行了预测,亦得到了较满意的结果,将另外撰文报道。总体建模及随机抽样建模结果(Y_{model} 、 Y_{pred})均与文献结果相近,但计算方法较文献简便许多,可望在蛋白质、多肽及核酸的QSAR研究中得到潜在应用。

3 结 论

论文在较为详细地综述定量构效关系研究进展以及本实验室提出的烷烃分子距边矢量基础上,基于点电荷作用原理,提出了电边矢量,仅从一级结构出发对寡肽进行了结构表征,以此为基础对血管紧张素转化酶、血管舒缓激肽增效剂等寡肽进行了定量结构活性相关研究,取得了与文献接近或更好的效果,同时从多元线性回归分析结果中获取了一些关于寡肽空间结构的信息,并且提出的电边矢量的计算比文献更为简便、更易获得。对于多肽,由于原子数目很多,如用基于原子为基础而构建的拓扑指数来进行构效关系的研究,

有其繁琐之处。于是又提出了以多肽链中氨基酸侧链为基础的电边矢量,使计算更为简便,并对抗原肽的活性进行了定量建模与预测,取得了良好的效果,向生物活性分子特别是大分子如蛋白质结构预测方面迈进了一步。

致谢

绿色化学与药物设计研究室及化学生物学与分子药理学研究室李声时教授、刘树深教授、周原副教授、杨胜喜工程师、廖春阳工程师、张梦军讲师,应用化学98-2班王远强、章仁辉、邹竹惠、吴世容、97-2班叶楠学士等提供有关帮助或协助。谨致谢忱。

参考文献:

- [1] PLACA J. Human genome - Development of Energy on the map[J]. Nature, 1986, 321: 371-373.
- [2] LUO YU, JIANG XULIANG, LAI LUHUA. Modeling protein backbone structure based on C_{α} guiding coordinates[J]. Protein Engineering, 1992, 5: 147-152.
- [3] ANFINSEN C B, HABER E, SELA M, et al. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain[J]. Proc Natl Acad Sci USA, 1961, 47: 1309-1318.
- [4] MARTIN J, MAYHEW M, LATTGER T, et al. The reaction cycle of GroEL and GroES in chaperonin - assisted protein folding[J]. Nature, 1993, 366: 228-231.
- [5] BOREA P A, SANTO G P, SALVADORI S, et al. Opioid peptides. Pharmacological activity and lipophilic character of dermorphin oligopeptides[J]. Farmaco. Ed. Sci. 1983, 38: 521-526.
- [6] WALLER C L, MARSHALL G R. Three dimensional quantitative structure activity relationship of angiotension - converting enzyme and thermolysin inhibitors. II. A comparison of CoMFA models incorporating molecular orbital fields and desolvation free energies based on active - analog and complementary - receptor field alignment[J]. J Med Chem, 1993, 36: 2390-2403.
- [7] RAYCHAUDHURY C, BANERJEE A, BAG P, et al. Topological shape and size of peptides: Identification of potential allele specific helper T cell antigenic sites [J]. J. Chem. Inf. Comput. Sci. 1999, 39: 248-254.
- [8] LIU SS, LI ZL. Approach to estimation and prediction for normal boiling points of alkanes based on a molecular distance - edge vector (MDE) [J]. J Chem Inf Comput Sci, 1998, 38(3): 387-394.

On Structural Parameterization and Functional Prediction of Oligopeptide Sequences with Tachykinin Activity through Quantitative Sequence - Activity Models by Molecular Electronegativity - Distance Vector

CHOU Liang-jia¹, LI Bo¹, LI Jin-wei¹, WANG Xiao-yan^{1,2}, LI Zhi-liang¹

(1. College of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400044, China;
2. College of Bioengineering, Chongqing University, Chongqing 400044, China)

Abstract: Only from the primary structure of peptides, a new set of descriptors called the molecular molecule electronegativity edge vector (VMED Vmed) which applied to describe the peptide structure is proposed, based on the distance between atom pairs and the electronegativity of each atom. Several models of quantitative structure activity relationship (QSAR) are proposed to various oligopeptides including 58 dipeptides with angiotensin converting enzyme inhibition and 31 pentapeptides by multiple linear regression method. In order to explain the ability to characterize the molecular structure of polypeptides, a molecular modeling investigation on QSAR is further carried out for functional prediction of heptapeptide sequences with tachykinin activity through quantitative sequence - activity models (qsams) by molecular electronegativity - edge vector (Vmed). The results show that VMEE exhibits both of excellent structural selectivity and good activity estimation. Besides, because VMED is easy to calculate and able to work, it will be very useful in structural characterization and activity prediction of biological molecules.

Key words: quantitative structure - activity relationship; molecular electronegativity - edge vector; bioactive oligopeptide; quantitative sequence - activity models

(编辑 张小强)