第 27 卷 第 2 期
2006 年 5 月

质 谱 学 报
Journal of Chinese Mass Spectrometry Society

Vol. 27 No. 2
May 2006

单质子化蛋白质肽段结构表达及其离子迁移谱碰撞截面定量预测

曾 晖[1,2],李志良[1,2,*],赵 娜[2,3],张巧霞[1,2],梅 虎[2,3],
周 原[1,2,3],杨善彬[2,3],李经纬[1,2]

(1. 重庆大学化学化工学院,重庆 400044; 2. 生物工程教育部与重庆市重点实验室,重庆 400044;
3. 重庆大学生物工程学院,重庆 400044)

摘要:基于分子二维图形特征提出一种新型结构参数化方法:分子电性作用矢量(MEI);应用于考察 107 个单质子化肽段样本集结构表征及离子迁移谱碰撞截面模拟和预测及严格检验,所得 3 个模型回归建模及留一法交叉验证复相关系数(R 和 Q)分别为:$R=R_{cu}=0.983$, 0.981, 0.980 和 $Q=R_{cv}=0.979$, 0.979, 0.978。MEI 对复杂有机分子结构表达准确且性质预测良好。

关键词:分子电性作用矢量;离子迁移谱;碰撞截面;定量构效关系

中图分类号:O657.63; Q516 **文献标识码**:A **文章编号**:1004-2997(2006)02-84-06

Molecular Structural Characteristics of Singly Protonated Peptides for Proteins and Quantitative Prediction of Collision Cross Section for Ion Mobility Spectrometry

ZENG Hui[1,2], LI Zhi-liang[1,2,*], ZHAO Na[2,3], ZHANG Qiao-xia[2,3],
MEI Hu[2,3], ZHOU Yuan[1,2], YANG Shan-bin[2,3], LI Jing-wei[1,2]

(1. *College of Chemistry and Chemical Engineering, Chongqing University, Chongqing* 400044, *China*;
2. *Key Laboratory of Biomedical Engineering of Educational Ministry and Chongqing Municipality, Chongqing* 400044, *China*;
3. *College of Bioengineering, Chongqing University, Chongqing* 400044, *China*)

Abstract: Molecular structural characteristics of singly protonated peptides for proteins are primarily investigated to perform quantitative prediction of collision cross section for ion mobility spectrometry. Based on two-dimensional topological characterization, a novel description vector called molecular electronegativity interaction vector (MEI) is proposed to express the structural characterization of molecule. Estimation and prediction of collision cross section for ion mobility spectrometry (IMS) of 107 singly protonated peptides for proteins were successfully done through MEI description vector. A good model is strictly es-

收稿日期:2005-07-08;**修回日期**:2006-01-09

基金项目:霍英东基金(1998),国家春晖计划教育部启动基金(1999-1-4/38),湖南大学化学生物传感与计量学国家重点实验室(2005-12),重庆大学创新基金(03-5-6+04-9-1)资助

作者简介:曾 晖(1980~),女(汉),湖南人,硕士研究生,分析化学专业。E-mail:zenghui1106@yahoo.com.cn

* **通讯作者**:李志良(1962~),男(汉),湖南人,博士生导师,教授,从事系统化学生物学及其组学信息研究。E-mail:zlli2662@163.com

tablished by multiple linear regression (MLR) with both cumulative multiple correlation coefficient R_{cum} and leave-one-out cross-validation Q_{LOO} are 0.983 and 0.979, respectively. From the obtained results, it is suggested that MEI is an excellent vectorial descriptor with both good structural selectivity and high property correlation, and quite suitable for quantitative structure property relationship for spectroscopy (QSPR/QSSR).

Key words: molecular electronegativity interaction vector (MEI); ion mobility spectrometry (IMS); collision cross section (CCS); quantitative structure-property relationship (QSPR); quantitative structure-spectrum relationship (QSSR)

近年来电喷雾(ESI)[1]、基质辅助激光解析(MALDI)[2]等新型离子化技术使得生物大分子被用于离子迁移谱(IMS)[3-4]检测成为可能,在此基础上将高灵敏度常压分析技术 IMS 与其它分析仪器联用(如 GC-IMS, IMS-TOFMS 等)可更为有效地扩展 IMS 在分子生物领域应用范围[5-6],包括广泛应用于反恐、航空、海关、战场、犯罪现场的痕量物质检测。碰撞截面(Ω, CCS)在 IMS 中是一个衡量离子结构特征的重要参数,它可由 Mason 等[7]提出的公式计算。式中涉及 z 为离子所带电荷数,e 为单位电荷电量 $1.602\,189\times10^{-19}$, k_b 是 Boltzmann 常数 $1.380\,62\times10^{-23}\cdot K^{-1}$, m 和 M 分别代表离子及迁移气体分子质量, P 和 T 为实验压力和温度, E 为所施加电场强度, t_D 是离子迁移时间, N 为迁移气体密度, L 为迁移管的长度。因此只要测量出 E, L, P, T, t_D 就可以通过该式精确计算被测离子 Ω 大小。然而由于实验条件以及检测技术的限制,大规模测量各类离子碰撞截面还难以实现,本文试图将基于统计学方法发展起来的定量构性/效关系(QSPR/QSAR)[8-9]应用于碰撞截面预测中,以期得到具有一定实用价值的定量模型。考虑到碰撞截面与被测物质结构特征有着密切关系,发展一种用于表征化学结构的电性拓扑指数:分子电性作用(MEI)矢量,并成功将其应用于 107 个单质子化肽段碰撞截面的建模与预测中。

1　原理和方法

1.1　基本概念

(1)原子类型(atomic type, AT)。按照元素周期表中"同族元素具有相似化学性质"的规律,将原子按族分类是较为简单且化学意义明确的方法。这里首先可将有机物中常见非金属元素原子分为 5 类,1 (IVA): C, Si; 2(IA): H; 3 (VA): N, P; 4(VIA): O, S; 5 (VIIA): F, Cl, Br, I。对于 5 类原子间的相互作用则可以得到 15 种不同的作用情况(1-1, 1-2, 1-3, 1-4, 1-5; 2-2, 2-3, 2-4, 2-5; 3-3, 3-4, 3-5; 4-4, 4-5; 5-5)。注意 MEI 并不限制其它原子加入,当分子中含有稀有原子时可基于按族分类方法定义将上述 5 种类型进一步扩充。由于天然氨基酸中不含第 5 类卤素原子故实际上只有 4 种类型,故计算氨基酸描述子时无第五类卤素原子作用项(即 10 项)相互作用。

(2)原子相对电性(atomic relative electronegativity, ARE)。通常将分子中的原子看作是没有体积且带有电荷的质点,原子之间通过化学键传递电性而发生相互作用。此外,参照定义电负性是指不同原子吸引电子能力的相对大小,其在一定程度上反映了分子内部电荷分布情况,这里使用鲍林电负性(Pauling's electronegativity, PE)[10]作为原子的电性标度,进一步规定计算中取原子的相对电性(ARE),即该原子相对于碳的电负性大小。对于各类原子的鲍林电负性[11]和其相对电性大小列于附表 1。

(3)原子相对键距(atomic relative bond-distance, ARB)。分子中的原子是通过各种类型化学键连接起来的一个整体,因此化学键成为原子相互作用最直接的传导介质。由于原子相互作用是随其连接距离的增加而迅速减小,因此可以将原子的作用距离看作是连接它们的最短键距而忽略其它连接路径。为达到数据统一目的,这里仿照原子相对电性处理办法将 C-C 单键键长视为 MEI 的标准键长,而其它化学键键长与其比值就为相对键长(RBL)。进一步将分子中原子相对键距(ARB)定义为连接两原子最短路径中所有化学键相对键长(RBL)之和。有机化合物中一些常见化学键长[12]和相对键长列于附表 2。

1.2 分子结构表达方法

分子体系中原子是以带有一定数量电荷的微观质点形式存在，电荷之间相互作用构成分子内部结构特征支配方式，而这种分子内部微观作用力又以相应的物化性质在物质宏观状态上表现出来。描述点电荷相互作用基本公式为库仑(Coulomb)定律，利用该定律可表达出分子内部电荷作用方式。本文将常见原子类型按元素周期表的族划分为 4 类，由于具有不同化学性质的各类原子之间相互作用所产生的效果有所差异，因此将各类作用情况视为不同的描述分量加以区分。最终得到 10 个 MEI 描述子分别由以下各类原子作用产生：1-1，1-2，1-3，1-4，2-2，2-3，2-4，3-3，3-4，4-4(表 1)。计算公式如下：

$$V_{kl}(\mathrm{E}) = \sum_{i\in k,\, j\in l}(\mathrm{e}^2/4\pi\varepsilon_0)(\mathrm{ARE}_i \cdot \mathrm{ARE}_j/(\sum_{i\to j}, \mathrm{RBL})^2) = \sum_{i\in k, j\in l}(\mathrm{e}^2/4\pi\varepsilon_0)(\mathrm{ARE}_i \cdot \mathrm{ARE}_j/\mathrm{ARB}_{ij}^2),\ (1\leqslant k\leqslant 5,\ k\leqslant l\leqslant 5) \quad (1)$$

这里 k 与 l 代表原子所属种类；ARE 为原子相对电性大小；ARB_{ij} 是指从 i 原子到 j 原子所经历最短路径上各个化学键相对键长之和。

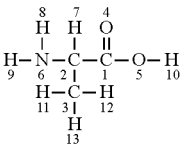

图 1 丙氨酸分子结构(图中数字为原子编号)

Fig. 1 Structure and molecular skeleton of alanine with numbered atoms

1.3 计算实例

以简单的丙氨酸(alanine，A)为例扼要说明。有关分子结构及原子编号见图 1，含有 C，H，N，O(S)四种原子；该分子 10 个非零描述子计算如下：对于碳-碳原子相互作用包括 1-2 号、1-3 号和 2-3 号原子作用情况，可以表达为下式：$\nu_{CC}=(1.000\,0\times1.000/1.000\,2)\times2+(1.000\,0\times1.000/2.000\,2)=2.250\,0$。同理可以计算剩下 9 个 MEI 描述子：$\nu_{CH}=10.293\,7$，$\nu_{CN}=1.932\,7$，$\nu_{CO}=4.826\,9$，$\nu_{HH}=2.838\,9$，$\nu_{HN}=5.484\,7$，$\nu_{HO}=4.426\,6$，$\nu_{NN}=0.000\,0$(该项在其它氨基酸或肽中可能不为零)，$\nu_{NO}=0.406\,7$，$\nu_{OO}=0.614\,6$。

2 实验部分

2.1 程序设计方案选择

有关方框图从略。这里仅仅给出流程图：1)划分原子类型(C[IVA]、H[IA]、N[VA]、O[包括 S，VIA])，将分子中原子类型及连接方式以字符形式存储在适当文本文件中；→2)获得二维拓扑结构(2D Topo)，程序读入该文本文件并视分子为顶点被不同原子染色的带权无向图；→3)引入对应参数(元素电负性、原子相对电性、化学键长、原子相对键距等)，采用 Floyd 算法找寻分子中所有原子间最短路径；→4)计算分子电负性作用矢量描述子 MEI(利用公式 1，MEI-tool.exe)进而同时计算该分子 10 个 MEI 描述子。

2.2 程序实现过程说明

使用 C 语言开发应用程序 MEI-tool.exe 来实现 MEI 描述子计算。用户将分子中原子类型及连接方式以字符形式存储在 MEI.txt 命名的文本文件中，程序读入该文件并视分子二维拓扑结构为一个顶点被不同原子染色的带权无向图，进而采用 Floyd 算法找寻分子中所有原子间最短路径同时利用(1)式计算该分子 10 或 15 个 MEI 描述子。为了简化类似于肽(或核酸)分子初始结构输入并保证其正确性，我们同时编写了序列格式转化程序 STM-tool.exe 以作为 MEI-tool.exe 方式的配套工具，此时将只用以单字母形式表示肽(或核酸)一级结构并由 STM-tool.exe 自动转化生成 MEI-tool.exe 的输入文件。

2.3 实验观测数据获取

从 Jurs PC 等[10]的报告中取得 107 个单质子化肽一级序列及所对应实验观测碰撞截面作为总体样本集(附表 3)，肽段长度从 5-10 共跨越六个等级且都是以赖氨酸(lysine，K)为 C 端终止残基以确保正电荷被固定于此位点。该组样本均从胰蛋白酶类物质产生，其离子碰撞截面使用 IMS-TOFMS 技术测得。首先通过 STM-tool.exe 和 MEI-tool.exe 计算得到所有肽 MEI 描述子，删除其中全零项后每个肽可由 10 个描述子表征(附表 3)。

3 结果及分析

通过多元线性回归(multiple linear regression，MLR)技术将其与碰撞截面建立起 10 参数回归模型 M1(Model 1)如下(回归系数置信区间对应的置信度为 95%)：

$$\Omega=(73.994\pm23.577)-(1.342\pm1.339)\nu_{CC}+(2.412\pm1.488)\nu_{CH}-(1.324\pm1.239)\nu_{CN}-(0.260\pm1.503)\nu_{CO}-(4.812\pm3.697)\nu_{HH}+(1.743\pm1.555)\nu_{HN}+(1.444\pm1.082)\nu_{HO}+(22.934\pm37.967)\nu_{NN}-(4.754\pm7.932)\nu_{NO}-(1.228\pm6.726)\nu_{OO}$$

$$(n=107,\ m=10,R_{cum}=0.983,\mathrm{RMS}=5.944,F=293.256,Q_{LOO}=0.979,\mathrm{RMS}_{CV}=6.575;F_{CV}=293.256) \quad (2)$$

式中下标 O 原子同时代表了 S 原子。对该模型使用留一法(leave-one-out, LOO)进行交叉校验(cross-validation, CV)[14]检测,即依次将训练集中每个样本取出,并由所剩下样本集(包含 107 个样本)建立多元线性回归模型对所取出样本进行预测。最终获得模型 M1 相关统计量见表 2,从中可看出 MEI 描述子能很好地与碰撞截面呈线性相关。由于多元线性回归方程可能出现多重共线性(multicollinearity)以及变量不显著等病态特征,因此需要对模型 M1 进行更为深入的讨论。采用 SPSS 13.0 统计学软件包对 M1 进行回归诊断,分别计算出 10 个变量的 t 统计值(t)、显著水平(α)、偏相关系数(P)以及方差膨胀因子(variance inflation factor, VIF$=1/(1-r_i^2)$, 其中 r_i 代表模型中第 i 个变量对其它变量的相关系数)列于表 3 当中,通过分析该表可发现模型 M1 似具有一定程度多重共线性(某些变量 VIF 值偏高),而且并不是所有变量都表现出显著特征(部分变量 t 值落在 $-2\leqslant t\leqslant 2$ 范围)。这可解释为 MEI 计算方法统一性使得不同描述子之间可能具有一定信息重叠。

将模型对 107 个肽的碰撞截面估计结果和交叉检验预测值与实验观测值相关情况分别绘于图 2a,2b 中。同时建立其残差分布散点图(图 3a),从该图可看到 107 个肽几乎没异常点,仅有 3 个样本残差值略微超出正负两倍均方根误差(Root Mean Square error,RMS)(虚线以外),但也远远小于三倍均方根误差(通常确定异常样本的界限)。使用 Cook 距离对中心化杠杆值(centered leverage,CL)作图证实上述结论(图 3b),其表现为所有样本点都分布在该图左下角一个很小的范围以内(Cook$<$0.1, CL$<$0.15)。另外由于 Tropsha 等[15-17]最近研究结果显示仅以 Q_{LOO} 作为模型稳定性能的检测标准并不能有效说明问题,为严格验证模型对外部样本的预测能力,从总体样本集中抽取 9 个分子作为测试集,并由剩下 98 个样本作为训练集建立 10 参数回归模型对其进行预测。由于 107 个肽的序列长度从 5 到 10 共分六个等级,为了使得测试结果更具可信度这里抽样规则满足以下 4 点要求:①每一等级长度的肽都随机选取 2 个样本进入测试集;②被选取的样本其序列上每一个位点的氨基酸残基类型在训练集同等级序列的该位置上必须出现过(限制自变量不能为外插值);③被选取的样本碰撞截面在同等级序列中不能为最大或最小值(限制因变量不能为外插值);④考虑到较长肽段可能更难表达和预测故在最长等级样本中多选择了一个肽以补足 9 个测试集样本数。最终选定测试集化合物在附表 2 中标以"*"标

表 1　模型 M1 的回归诊断结果

Table 1　Results of regressive diagnosis for model 1

变量 Variables	t 统计值 t	显著水平 α	偏相关系数 P	方差膨胀因子 VIF
constant	6.225	0.000	-	-
ν_{CC}	−1.986	0.050	−0.193	288.677
ν_{CH}	3.215	0.002	0.303	1900.894
ν_{CN}	−2.121	0.036	−0.206	90.663
ν_{CO}	−0.343	0.732	−0.034	215.814
ν_{HH}	−2.581	0.011	−0.248	800.788
ν_{HN}	2.223	0.028	0.215	138.812
ν_{HO}	2.646	0.009	0.253	62.501
ν_{NN}	1.198	0.234	0.118	251.694
ν_{NO}	−1.189	0.237	−0.117	236.706
ν_{OO}	−0.362	0.718	−0.036	44.180

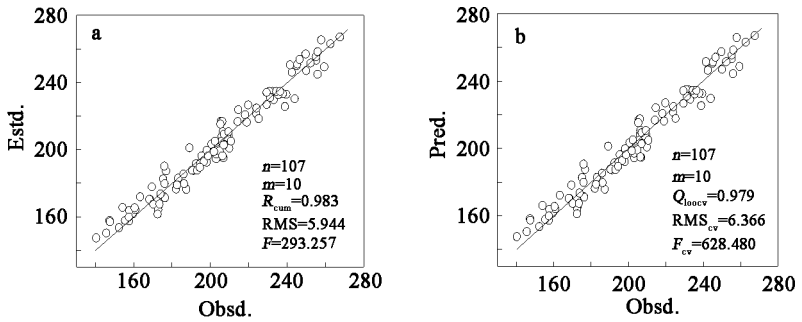

图 2 模型对 107 个肽碰撞截面实验观测值与(a)估计值及(b)交叉检验预测值相关图

Fig. 2 (a) Plot of estd. and (b) pred(LOO). vs. obsd. collision cross section of 107 singly protonated peptides

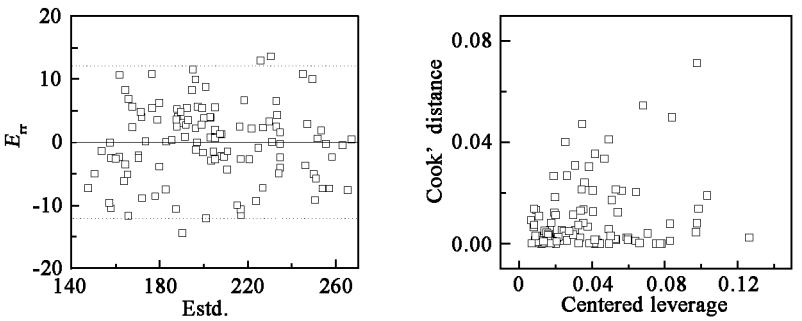

图 3 (a)模型对 107 个肽碰撞截面估计值残差分布散点图(虚线处为两倍均方根误差);
(b)样本 Cook 距离对 Centered Leverage 值作图

Fig. 3 (a) Distribution of residuals (dashed is double root mean square error);
(b) Plot of Cook's distance. vs. Centered Leverage values

记(由于总体样本多样性有限使得其中两个分子没满足条件②)。包含 98 个肽化合物训练集所建 10 参数回归模型相关统计量为:$R_{cum}=0.978$, RMS = 6.420, $F=520.026$, $Q_{LOO}=0.976$, $RMS_{CV}=6.749$, $F_{CV}=468.141$。模型对测试集 9 个样本预测值对实验值回归,其统计结果为:$R_{ext}=0.997$, $RMS_{ext}=2.761$。由此可见使用 MEI 描述子建立肽离子碰撞截面预测模型具很好稳定性能及泛化能力。

4 结 语

一种优良分子结构描述子须具较强的特征分辨率和良好的性质相关性,同时还应满足操作简便及计算迅速等多方面要求。在该思想指引下从分子图形二维拓扑不变量出发提出用于表达有机物结构分子电性作用矢量(MEI)。利用该矢量对 107 个肽化合物样本进行表征并与其 IMS 碰撞截面建立简单线性模型,取得了较为满意的结果。由于 MEI 从电性特征出发将原子按族分类,该法同样能够适用于包含大量杂原子的药物及生物分子结构表征、性质和活性预测。

致谢:感谢药物设计研究室李眚励、李至真、袁晓燕教授、周鹏硕士、李容副教授、张梦军讲师、林红卫、杨胜喜工程师、刘振德硕士、兰玉坤硕士、邹竹惠学士等提供有关软件、文献以及技术上面的帮助;化学生物学与分子药物学研究室李晟和、李声时教授、黄莺副教授、熊清博士、孙立力博士、聂金媛硕士、吴世容、李根容硕士和叶楠学士等,谨致谢意。

参考文献:

[1] Fenn J B, Mann M, Meng C K, et al. Electrospray Ionization for Mass Spectrometry of Large Biomolecules[J]. Science, 1989, 246: 64-71.

[2] Karas M, Hillenkamp P. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10 000 Daltons[J]. Anal Chem, 1988, 60: 2 299-2 301.

[3] Hill H H, Siems W F, Louis R H, et al. Ion Mobility Spectrometry. Anal Chem, 1990, 62:

1 201A-1 209A.

[4] Myung S, Lee Y J, Moon M H, et al. Development of High-Sensitivity Ion Trap Ion Mobility Spectrometry Time-of-flight Techniques: A High-throughput Nano-LC-IMS-TOF Separation of Peptides Arising from a Drosophila Protein Extract [J]. Anal Chem, 2003, 75: 5 137-5 145.

[5] Beegle L W, Kanik I, Matz L, et al. Electrospray Ionization High-Resolution Ion Mobility Spectrometry for the Detection of Organic Compounds, 1. Amino Acids[J]. Anal Chem. 2001, 73: 3 028-3 034.

[6] Revercomb H E, Mason E A. Theory of Plasma Chromatography/Gaseous Electrophoresis. Anal Chem, 1975, 47: 970-983.

[7] Katritzky A R, Maran U, Lobanov V S, et al. Structurally Diverse Quantitative Structure-Property Relationship Correlations of Technologically Relevant Physical Properties [J]. J Chem Inf Comput Sci, 2000, 40: 1-18.

[8] Hansch C; Hoekman D, Leo A, et al. Chembioinformatics: Comparative QSAR at the Interface between Chemistry and Biology[J]. Chem Rev, 2002, 102: 783-812.

[9] Pauling L. The Nature of Chemical Bond IV. Energy of Single Bonds and the Relative Electronegativity of Atoms[J]. J Am Chem Soc, 1932, 54: 3 570-3 582.

[10] Mosier P D, Counterman A E, Jurs P C, et al. Prediction of Peptide ion Collision Cross Sections from Topological Molecular Structure and Amino Acid Parameters [J]. Anal Chem, 2002, 74: 1 360-1 370.

[11] Wold S. Cross-validation Estimation of the Number of Components in Factor and Principal Components Models[J]. Technometrics, 1978, 20: 897-903.

[12] Golbraikh A, Tropsha A. Beware of q^2 ![J]. J Mol Graphics Mod, 2002, 20: 269-276.

[13] Tropsha A, Gramatica P, Gombar V K. The Importance of being Earnest: Validation is the Absolute Essential for Successful Application and Inerpretation of QSPR Models[J]. QSAR Comb Sci, 2003, 22: 69-77.

[14] Gramatica P, Pilutti P, Papa E. Validated QSAR Prediction of OH Tropospheric Degradation of VOCs: Splitting into Training-test Sets and Consensus Modeling[J]. J Chem Inf Comput Sci, 2004, 44: 1 794-1 802.

(上接第 83 页)

[9] Carlsson A. The Dopamine and Glutamate Hypotheses of Schizophrenia - Therapeutic Implications of an Integrated View [J]. European Neuropsychopharmacology, 1995, 5(3): 160.

[10] Weinbach E C. The Effect of Pentachlorophenol on Oxidative Phosphorylation [J]. J Biol Chem, 1954, 210:545-550.

[11] **常　浩,金泰廙. 五氯酚的内分泌干扰作用研究进展**[J]. **环境与健康杂志**,2002,19(3):279-281.

[12] **万大娟,李顺义,舒月红,等. 超声波法提取土壤样品中** 1-2-4 **三氯苯和** GC **测定**[J]. **辽宁城乡环境科技**,2005,17(2):102-104.

[13] Becker R, Buge H G, Win T. Determination of Pentachlorophenol (PCP) in Waste Wood- method Comparison by a Collaborative Trial [J]. Chemosphere, 2002, 47: 1 001-1 006.

[14] Munch J W. Method 508.1 Gas Determination of Chlorinated Pesticides, Herbicides and Organohalides by Liquid-solid Extraction and Electron Capture Chromatography [Z]. U. S. Environmental Protection Agency, Environmental Monitoring and Support Laboratory, Cincinnati, Ohio 45268, 1995.

[15] Iwata H, Tanabe S, Sakai N, et al. Persistent Organochlorine Residues in Sediments From the Chukchi Sea, Bering Sea and Gulf of Alaska [J]. Oceanographic Literature Review, 2002, 47: 1 001-1 006.