新氨基酸拓扑描述子得分矢量及支持向量机用于肽定量序列-迁移预测

梁桂兆[①②③] 梅 虎[②③] 杨善彬[①②] 周 原[①②③] 周 鹏[①②] 李志良[①②*]

(① 重庆大学化学化工学院, 重庆 400030; ② 湖南大学化学生物传感与计量学国家重点实验室, 长沙 410082; ③ 重庆大学生物力学与组织工程教育部重点实验室, 重庆 400030. * 联系人, E-mail: zlli2662@163.com)

摘要 收集了 23 种氨基酸 1262 种结构信息, 经主成分分析得新的-氨基酸拓扑描述子得分(SATD)矢量. 将其用于 125 个不同长度肽的结构表征, 分别以支持向量机(SVM)和偏最小二乘(PLS)建立肽定量序列-迁移模型(QSMM). 结果表明, SATD 描述子所含信息量大, 易于操作, 能较好地表征 125 个肽结构. 与 PLS 相比, SVM 在对电泳迁移率建模预测中表现出较强的拟合能力和良好的预测能力.

关键词 氨基酸 **SATD** 电泳迁移率 **QSMM** **SVM**

肽的分离、纯化和分析对药物开发具有重要意义. 在各种分离与分析方法中, 毛细管电泳具高效、高速和自动化等优点. 尽管其在肽和蛋白质分离和分析中得到广泛应用, 但对特定肽的实验优化仍然费时和费力. 定量结构性质关系(QSPR)则为研究电泳迁移行为提供一个思路: 以分子物化特性与结构信息相关为前提, 将其用于肽定量电泳迁移预测. 结构表征和建模方法是QSPR研究的重要内容. 目前对肽类物质结构表征虽已有研究[1,2], 但因结构复杂, 主要集中于 2D方面. 在此基础上, 提出一种新氨基酸描述子即氨基酸拓扑描述子得分矢量, 将其用于 125 肽序列表征. 多元线性回归、主成分回归和偏最小二乘(PLS)可用于线性建模, 二次PLS和神经网络等可用于非线性建模. 近年来, 支持向量机(SVM)[3]因具有显著的优点引起研究者的重视, 其基于结构风险最小化理论, 寻找一最优界面实现最小泛化误差, 能较好地解决非线性、高维数和局部最小等问题. 毛细管区带电泳因其迁移率同分子所带电荷与其半径比值成正比, 故具简单化和可预测性, 这种带电分子特性与迁移之间的简单关系为以定量序列-迁移模型(QSMM)预测肽类分子迁移方式提供可能, 故以SVM回归建立QSMM, 对了解肽电泳迁移规律具有一定的参考意义.

1 原理与方法

1.1 氨基酸拓扑描述子得分(SATD)

氨基酸构成及其物化性质参数与拓扑结构对肽生物功能至关重要, 故收集 23 种天然氨基酸的 1262 种性质参数: 0D参数包括 31 种构成描述子; 1D参数 69 种, 包括官能团数目、原子中心碎片和分子特性等; 2D参数 533 种, 包括本室提出的分子电性作用矢量, 分子电距矢量和全息分子电距矢量, 拓扑, 拓扑电荷指数, 运转和路径数目, 边缘邻接指数, Burden特征值, 自相关, 连接性指数, 信息指数及特征值指数等; 3D参数 629 种包括Randic分子剖面, 几何描述子, RDF, MoRSE, WHIMs和GETAWAY等[4~13]及本实验室提出的分子电性作用矢量、分子电距矢量和全息分子电距矢量[14~18]. 用主成分分析(PCA)[19]压缩变量数目, 最终得到解释原始变量矩阵(23×1262)方差95.12%的 13 个主成分得分矩阵(23×13)(附表 1)[1)], 代替原始变量矩阵. 从 13 个主成分载荷方程中的各变量系数可分析各变量对每个主成分的贡献大小. 对第一主成分正贡献较大的相关变量包括非氢原子数, Randic标识数, Balaban标识数及Randic连接性指数等; 负贡献较大的主要是特征值指数中原子质量、范德华及电负性权重距离矩阵平均特征矢量系数总和等. 第二主成分主要涉及 3D-MoRSE描述子中原子极化度及Sanderson电负性参数对应权重等. 第三主成分主要反映分子特性、官能团数目及特征值指数中基于碎片极性表面积、Moriguchi与Ghose-Crippen辛醇-水分配系数及氢键接受原子(N, O及F)数目等信息. 第四主成分主要涉及WHIM和GETAWAY 描述子中原子质量、极化度、范德华体积、电负性及电

1) 所有附表限于篇幅不拟发表, 可向作者函索

拓扑状态等变量. 对第五主成分影响较显著的是原子中心碎片 CH_3X(X 代表杂原子, 下同)数目及 3D-MoRSE 描述子中原子质量等信息. 对第六主成分正贡献较大的是拓扑描述子中 N 与 S 及 O 与 S 之间拓扑距离的总和, 几何描述子中 N 与 S 及 O 与 S 之间的几何距离总和; 负贡献较大的是几何描述子中芳香指数谐波振荡器模型(HOMA)及 HOMA 总量等信息. 第七主成分主要提炼分子中五元环数目、环一级 $C(sp^3)$数目及环二级 $C(sp^3)$数目, 3D-MoRSE 与 WHIM 描述子中原子质量及 Sanderson 电负性等信息. 第八主成分主要反映边缘邻接指数中边缘数目、偶极矩与共振积分加权矩阵信息及 GETAWAY 描述子中原子质量和极化度等. 第九主成分主要体现边缘邻接指数描述子中边缘数目, 3D-MoRSE 中原子质量和范德华体积, 自相关描述子中 Sanderson 原子电负性和范德华体积, 原子中心碎片 $H\text{-}C^3(sp3)$或 $H\text{-}C^2(sp2)$或 $H\text{-}C^3(sp2)$或 $H\text{-}C^3(sp)$数目, R--CH—X, R--CR--X 及 X--CH--X 数目(--表示与吡咯环中 C—N 键类似的芳香单键)等变量. 对第十主成分正贡献较大的是自相关描述子中 Sanderson 原子电负性和极化度, 3D-MoRSE 及 GETAWAY 中原子质量等信息, 负贡献较大的是自相关描述子中原子质量, 边缘邻接指数描述子中偶极矩与共振积分加权矩阵等变量. 第十一主成分中正载荷较大的是 GETAWAY 中原子质量、电负性、范德华体积及极化度等变量, 负载荷较大的是原子中心碎片 X—C(═X)—X(—表示与苯环中的芳香键类似的键或硝基中 N—O 之间的键)及醇数目等变量. 第十二主成分主要反映 3D-MoRSE 与自相关描述子中加权范德华体积, WHIM 描述子中原子质量、电负性及范德华体积, 自相关描述子中原子极化度等变量. 第十三主成分主要体现自相关、GETAWAY 及 3D-MoRSE 描述子中与原子质量和电负性等信息. 经上述分析发现, 此 13 个主成分主要提炼的是分子重量、体积、原子电性、构成及极化度等信息. 为简便, 将相应得分矢量命名为氨基酸拓扑描述子得分(scores of amino acid topological descriptors, SATD), 每个残基可用 13 个 SATD 变量表达, 对具 n 个残基的肽, 其序列可用 $13\times n$ 个 SATD 变量串联表征. 1262 种参数由 Dragon 软件(http://www.talete.mi.it/)及自编软件计算得到.

1.2 序列表征与自交叉协方差处理

以SATD表征长度为n的肽将产生 $13\times n$个变量, 若其残基数不同, 表征后将得不同数目变量. 为使其所有表征变量数目一致, 以自交叉协方差(ACCs)[15]对每个肽的原始表达变量数目做一致化处理. 该法考虑到肽链不同位点氨基酸参数之间的交互效应, 可最大程度降低信息损失, 具操作简便和快捷等优点. 不同长度的肽链经ACCs处理后变量数目皆为$13^2\times l$个(l为步长, 其小于最短肽链长度). ACCs程序由C语言编写实现.

1.3 变量挑选与模型验证

为降低模型复杂度, 提高稳定性、可解释性及预测能力, 建模前用遗传算法(GA)-PLS方法剔除变量中与活性相关性较小的信息. 其包括 5 个步骤[16]: (1) 基因编码与群体初始化; (2) 基因评价; (3)基因选择; (4) 遗传操作包括杂交、变异和复制; (5)反复执行(2)~(4)步骤, 直到达到终止条件, 选择最佳个体作为结果. 模型预测能力由适应度函数 Q^2_{cv} 评价[17](文献中为q^2, 对内部样本预测能力验证用 Q^2_{cv} 表示). 一般地, Q^2_{cv} 越大, 模型预测能力越强. 但外部验证必不可少, 可用 Q^2_{ext} 评价[16]. D-最优设计是一种将Fisher信息矩阵$|X'X|$行列式最大化的实验设计方法, 此处用来划分训练集和测试集. 其中X指自变量或自变量与因变量矩阵. 使$|X'X|$占据整个数据点空间, 这些点构成训练集, 没有被选中的点视为测试集. 将 125 个样本划分为 100/25, 用 100 个肽建立QSMM并进行内部验证, 用 25 个肽验证模型外部预测能力. GA-PLS和D-最优程序均由软件Matlab7.0 完成.

1.4 PLS 建模

PLS主要适于建立多自变量对多因变量线性回归建模, 同时适于样本数目小于变量数目情况下的回归, 可避免变量相关性危害, 具有回归建模、PCA和典型相关等优点[18]. PLS回归由软件Matlab7.0 完成.

1.5 SVM 建模

SVM 解决线性回归问题就是求一超平面, 使所有样本点到超平面的距离最小. 对非线性问题, 首先经一个非线性映射 Φ, 将样本映射到一个高维特征空间, 然后用线性方法来解决. 高维映射经核函数: $K(x, x_i)=\Phi(x)\cdot\Phi(x_i)$实现. 因 SVM 引入核函数, 故可有效避免维数灾难及计算复杂性等问题. 目前常用核函数主要有线性(linear)核: $K(x, x_i)=x\cdot x_i$; 多项式

(polynomial)核: $K(x, x_i)=(\alpha_1 x \cdot x_i+\alpha_2)^p$; RBF 核: $K(x, x_i)=\exp(-\gamma \| x-x_i \|^2)$; sigmoid 核: $K(x, x_i) = \tanh(\alpha_1 x \cdot x_i+\alpha_2)$. 为与线性 PLS 建模方法比较, 选择非线性的 RBF 核 SVM 建立 QSMM. 参数选择对 SVM 回归成败至关重要. 其中, 惩罚系数 C 是一个调控参数, 用于调节最大间隔和最小化训练误差之间的平衡. 不敏感损失函数中 ε的优化主要依靠未知数据集中噪声类型, ε不敏感性阻止整个训练集超出边界条件, γ控制 SVM 的泛化能力. 因对参数选择没有严格标准, 本研究借鉴正交实验设计思想, 根据 Q^2_{ext} 确定 C, ε和γ的优化值. SVM 回归由软件 Matlab7.0 实现.

2 结果与讨论

2.1 PLS 建模分析

所选 125 个肽来自Mehdi等 [19]的报告, 所含残基为 2~14 个, 用SATD表征ACCs处理后每个肽由 169 个变量表征. GA-PLS参数设置如下: 初始群体大小: 200; 最大遗传代数: 200; 收敛标准: 80%; 交叉频率: 50%; 变异概率: 0.5%; 从 10 个模型中(附表 2), 选择一包含 47 个变量的最优模型(模型 7). 用D-最优将整个样本集划分为 100(编号 1~100)个训练集样本和 25 个(编号 101~125)个测试集样本. 用所选 47 个变量以 PLS建模, 得 3 个显著主成分, 分别解释原始变量 52.9%, 18.4%和 10.0%的方差, 均方根误差(RMS)=3.766. 对 25 个测试集预测得 Q^2_{ext} =0.209, RMS_{ext}=6.887, 预测值表明, PLS模型拟合效果较好, 但对外部样本预测能力较差. PLS模型得分图(图 1)表明样本 27, 41, 47, 49, 56, 82, 17, 21, 30, 45 和 79 位于 Hotelling T^2 椭圆置信区间外, 为特异点, 但此 11 个样本在空间内呈两个区域分布, 说明样本点 27, 41, 47, 49, 56, 82 与样本点 17, 21, 30, 45, 79 的高维空间性质相差较远. 从PLS模型载荷图(图 2)可看出各变量与Y(迁移率)的相关性. 在第一主成分中变量v_{11} 和v_{43}的载荷较大(大于 0.2)与Y且呈正相关, v_3, v_{35}, v_{38}, v_{42} 和v_{47} 的载荷小于–0.2, 与Y呈负相关. 第二主成分中, 变量v_8, v_{19}, v_{27}, v_{44}和v_{46}对Y贡献较大, 并与Y呈正相关, 而v_{14}, v_{28}, v_{29}, v_{34}, v_{35}和v_{45}与Y呈负相关(载荷小于–0.2). 值得注意的是, 在两主成分中, 变量v_{35}载荷均较大, 且均与Y呈负相关.

2.2 SVM 建模分析

研究发现, 以 169 个原自变量为 SVM 的输入,

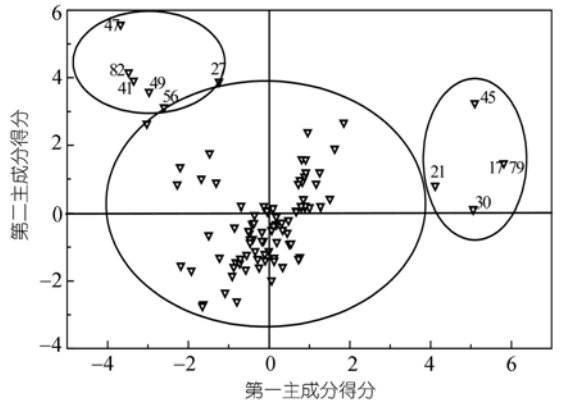


图 1 PLS 模型得分图

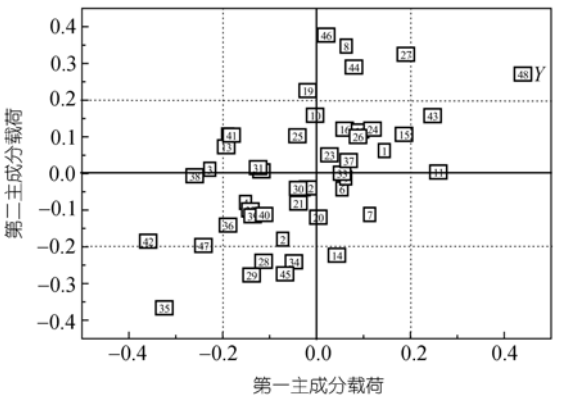


图 2 PLS 模型载荷图

建模及外部预测效果都较差, 故以 GA-PLS 所选 47 个变量为 SVM 的输入. 首先用单因素轮换法确定各因素的大致范围, 根据 Q^2_{ext} 随 C, ε 及 γ 变化图(图 3), 选择 C=200~500, ε= 0.4~0.7, γ= 0.014~0.056, 设计 3 因素 4 水平正交预测实验优化 C, ε和γ(附表 3), 从 16 次验证结果看, 当 C = 300, ε = 0.5, γ= 0.014 时具最大 Q^2_{ext} = 0.722. 故 SVM 回归参数为: C = 300, ε =0.5, γ=0.014(模型 6), 此时核函数为: $K(x, x_i) = \exp(-0.014 \| x-x_i \|^{1)})$. 在此条件下建模得 R^2=0.942, RMS = 2.087, 外部验证得 Q^2_{ext} =0.722, RMS_{ext}=4.079. 预测结果及相关统计量显示, SVM 模型拟合能力及外部预测能力都不同程度优于 PLS. 预测值与实验值回归(图 4)表明用 SVM 可较好拟合电泳迁移率与 SATD 的关系.

虽 SVM 建模结果及外部预测能力优于 PLS, 但亦出现一些离群值(图 4), 如 SVM 对 21 样本 RW 的预测值明显差于 PLS, 将其与 19 号样本 LW 和 20 号样本 PW 对比发现, 虽 L 和 P 分子半径小于 R, 但因

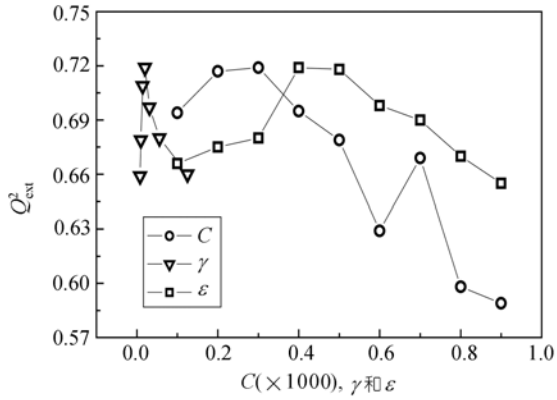


图 3 Q^2_{ext} 随 C, ε 及 γ 的变化图

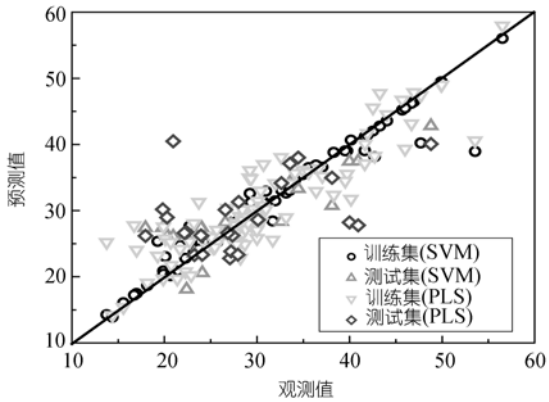


图 4 预测值与实验值回归图

L 和 P 皆带有非极性侧链, 其易结合相对较多电荷的作用, 这种作用可能“屏蔽”电泳迁移率因分子半径增大而减少的趋势, 故 RW 的观测值较 LW 和 PW 大, SVM 对其预测值偏低, 表明可能模型对此种“屏蔽”作用不敏感. 在 53 号分子 FGFG 中, G 是一很有特点的氨基酸, 其 α 碳原子上只有两个氢原子而无侧链, 分子半径相对较小, 同时, 与其他残基相互作用较小, 亦不产生任何空间位阻现象. 同等条件下, 电泳迁移率有增大趋势. 但因体积较大的 F 产生的空间位阻效应可能超过此趋势, 故推测 PLS 模型对这种效应识别不敏感导致对其预测值远大于观测值, 而 SVM 可较好地识别此效应, 故预测误差较小. 对 55 号样本 RFDS, 虽然此分子中 R, D 和 S 氨基酸是极性氨基酸, 可结合较多电荷, 使电泳迁移率呈变大趋势, 但因体积较大的 F“屏蔽”了此趋势, PLS 模型对这种“屏蔽”作用识别较敏感, 导致预测值远小于观测值. 对于 FFFF 样本, 对比第 6 号 FF 样本及第 107 号样本 FFF 发现, 随着 F 残基数目增加, 模型对其预测误差越来越大, 虽 F 可带有一定电荷, 但随体积较大的 F 残基数增多, 产生了所带电荷与其分子直径比值越来越小的变化, 推测 PLS 和 SVM 模型对此种变化识别相对不敏感, 造成预测值明显高于观测值.

3 结论

提出一新氨基酸描述子——SATD, 将其用于 125 个不同长度肽结构表征, 用 SVM 和 PLS 建立 QSMMs. 研究表明, SATD 描述子具操作简便、表达能力强和含信息量大等优点. 用 SVM 回归可获得优于 PLS 的建模结果, 其所建模型表现稳健且预测能力强. 但应看到, 因建模时所选参数较多且复杂, 要获得更成功的应用, 需对 SVM 原理和应用技巧进一步探讨. 本研究提出用正交设计方法挑选变量是一较好的尝试, 研究结果亦证明该方法的有效性. SVM 建模在肽 QSMM 研究中可进一步推广.

致谢 本工作为重庆市应用基础基金(批准号: 01-3-6)及湖南大学化学生物传感与计量学国家重点实验室基金资助项目.

参 考 文 献

1 Hellberg S, Sjöström M, Skagerberg B, et al. Peptide quantitative structure-activity relationships, a multivariate approach. J Med Chem, 1987, 30: 1126—1135[DOI]
2 Mei H, Zhou Y, Li S Z, et al. A new set of amino acid descriptors and its application in peptide QSARs. Biopolymers(Peptide Sci), 2005, 80(6): 775—786[DOI]
3 Cortes C, Vapnik V. Support vector networks. Machine Learning, 1995, 20: 273—293
4 Tdeschini R, Consonni V. Handbook of molecular descriptors. Weinheim: Wiley-VCH, 2000
5 Gilvez J, Garcia R, Salabert M T, et al. Charge indexes: new topological descriptors. J Chem Inf Comput Sci, 1994, 34: 520—525[DOI]
6 Rucker G, Rucker C. Counts of all walks as atomic and molecular descriptors. J Chem Inf Comput Sci, 1993, 33: 683—695[DOI]
7 Balaban A T, Ciubotariu D, Medeleanu M. Topological indices and real number vertex invariants based on graph eigenvalues or eigenvectors. J Chem Inf Comput Sci,1991, 31: 517—523[DOI]
8 Diudea M V, Horvath D, Graovac A. Molecular topology. 15. 3D distance matrices and related topological indices. J Chem Inf Comput Sci, 1995, 35: 129—135[DOI]
9 Randic M, Kleiner A F, DeAlba L M. Distance/distance matrices. J Chem Inf Comput Sci, 1994, 34: 277—286[DOI]
10 Schuur J H, Selzer P, Gasteiger J. The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. J Chem Inf Comput Sci, 1996, 36: 334—344[DOI]
11 Gasteiger J, Sadowski J, Schuur J, et al. Chemical information in 3D space. J Chem Inf Comput Sci, 1996, 36: 1030—1037[DOI]
12 Todeschini R, Gramatica P, Provenzani R, et al. Weighted holistic invariant molecular descriptors. Part 2. Theory development and

applications on modeling physicochemical properties of polyaromatic hydrocarbons. Chemon Intell Lab Syst, 1995, 27: 221—229[DOI]
13 Consonni V, Todeschini R, Pavan M. Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 1. Theory of the novel 3D molecular descriptors. J Chem Inf Comput Sci, 2002, 42: 682—692[DOI]
14 邓洪, 黄萍, 胡银玉, 等. 新分子距离边数矢量用于环境毒物多氯代二苯并呋喃的定量结构保留相关键模. 科学通报, 2005, 50(9): 875—879
15 梅虎, 梁桂兆, 周原, 等. 支持向量机用于定量构效关系建模的研究. 科学通报, 2005, 50(15): 1503—1506
16 Liu S S, Cai C C, Li Z. Approach to estimation and prediction for Normal Boiling Point (NBP) of alkanes based on a novel Molecular Distance-Edge (MDE) vector. J Chem Inf Comput Sci, 1998, 38, 387—394[DOI]
17 Li S Z, Fu B H, Liu S S, et al. On structural parameterization and molecular modeling of peptide analogues by Molecular Electronegativity Edge Vector (VMEE): Estimation and prediction for biological activity of pentapeptides. J Chin Chem Soc, 2001, 48: 937—944
18 Liu S, Yin C, Cai S, et al. A novel MHDV descriptor for dipeptide QSAR studies. J Chin Chem Soc, 2001, 48: 253—260
19 Liu S, Yin C, Cai S, et al. QSAR study of steroid benchmark and dipeptides based on MEDV-13. J Chem Inf Comput Sci, 2001, 41, 321—329[DOI]
20 Kim D, Lee I -B. Process monitoring based on probabilistic PCA. Chemon Intell Lab Syst, 2003, 67: 109—123[DOI]
21 Andersson P M, Sjöstrom M, Lundstedt T. Preprocessing peptide sequences for multivariate sequence-property analysis. Chemom Intell Lab Syst, 1998, 42: 41—50[DOI]
22 Tropsha A, Gramatica P, Gombar V K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci, 2003, 22: 69—77[DOI]
23 Liu H X, Zhang R S, Yao X J, et al. Prediction of the isoelectric point of an amino acid based on GA-PLS and SVMs. J Chem Inf Comput Sci, 2004, 44: 161—167[DOI]
24 Wold S, Sjöström M, Eriksson L. PLS-Regression: a basic tool of chemometrics. Chemom Intell Lab Syst, 2001, 58: 109—130[DOI]
25 Mehdi J-H, Shen Y, Hassanisadi M, et al. Prediction of electrophoretic mobilities of peptides in capillary zone electrophoresis by quantitative structure-mobility relationships using the offord model and artificial neural networks. Electrophoresis, 2005, 26: 1874—1885[DOI]

(2006-01-13 收稿, 2006-05-16 接受)