论著

文章编号:1000-5404(2013)13-1366-05

基于分组重量编码和特征选择技术预测外膜蛋白

南重汉[1], 邹凌云[2] (721006 陕西 宝鸡,解放军第3医院传染结核病防治所[1];400038 重庆,第三军医大学基础医学部微生物学教研室,重庆市微生物工程实验室[2])

[摘要] **目的** 建立新的机器学习模型,从蛋白质数据集或全基因组蛋白质序列中预测外膜蛋白。**方法** 采用分组重量编码和氨基酸组成计算蛋白质序列特征,采用F-score方法反向选择特征,采用支持向量机算法建立分类模型,在1 087条蛋白质序列构成的数据集上进行测试,评价预测模型的敏感性、特异性和预测精度,在多个细菌的全基因组蛋白质中预测外膜蛋白。**结果** 新的序列组合编码方法与SVM相结合,区分外膜蛋白和α螺旋跨膜蛋白、球状蛋白、非外膜蛋白的准确度分别达到94.7%、96.4%和94.6%,经特征选择之后,分类准确度分别提高到95.7%、96.9%和95.9%,且在基因组数据集中的预测结果与已知事实相符度高。**结论** 该方法预测准确度优于其他基于序列特征的预测方法,可用于在基因组序列中预测和筛选新的外膜蛋白。

[关键词] 外膜蛋白;分组重量编码;特征选择;支持向量机

[中图法分类号] Q51;Q755;Q78 [文献标志码] A

Prediction of outer membrane proteins with encoding based on grouped weight and feature selection techniques

Nan Chonghan[1], Zou Lingyun[2] ([1]Institute of Infectious TB Prevention, No. 3 Hospital of PLA, Baoji, Shaanxi Province, 721006; [2]Department of Microbiology, Chongqing Laboratory of Microbial Engineering, College of Basic Medical Sciences, Bioinformatics Center, Third Military Medical University, Chongqing, 400038, China)

[**Abstract**] **Objective** To establish a new machine learning model, and to predict outer membrane proteins (OMPs) from various protein datasets and genomic protein sequences using the methods with better accuracy than existing methods. **Methods** For protein sequences, combined features were constructed by encoding based on grouped weights and computing amino acid compositions, and the F-score technique was used to obtain backward selection subsets, which were then inputted to a classifier based on support vector machine algorithms (SVM). The classification model was tested on a dataset including 1 087 protein sequences for evaluation of sensitivity, specificity and overall accuracy, and was also trained and performed to predict OMPs in multiple bacterial genomes. **Results** Through combing the new sequence encoding methods with a SVM-based classifier, our model discriminated OMPs from α-helical transmembrane proteins, globular proteins and non-OMPs with cross-validated accuracy of 94.7%, 96.4% and 94.6%, respectively, and the accuracy severally increased to 95.7%, 96.9% and 95.9% after the feature selection. The whole-genomic prediction also fit the known facts reasonably. **Conclusion** The presented model shows better accuracy than other methods based on sequence features in the literature, and can be used for screening OMPs in genomic sequences.

[**Key words**] outer membrane protein; encoding based on grouped weight; feature selection; support vector machine

Supported by the Natural Science Foundation of Chongqing (2010BB5172) and the Innovative Foundation of Third Military Medical University (2010XQN05).
Corresponding author: Zou Lingyun, E-mail: zoulingyun1979@126.com

外膜蛋白(outer membrane proteins,OMPs)普遍存在于革兰阴性细菌、真核生物线粒体或者叶绿体的外膜[1],行使多种多样的生物学功能,如主动和被动运输(某些分子的运输通道)、酶促活性、细胞黏附和结

[基金项目] 重庆市自然科学基金(2010BB5172);第三军医大学科研创新基金(2010XQN05)
[通信作者] 邹凌云,电话:(023)68771932, E-mail: zoulingyun1979@126.com
[优先出版] http://www.cnki.net/kcms/detail/51.1095.R.20130227.1337.003.html (2013-02-27)

构锚定、抗原表位等。随着大量从基因组注释获得的蛋白质序列数据的产生,利用计算的方法从这些序列中预测可能的 OMPs,以及进一步揭示其跨膜区的结构成为人们关注的问题。现有的预测方法大多是利用氨基酸序列的组成、疏水性等特征[2],并且结合神经网络[3-4]、支持向量机(support vector machine, SVM)[5],K-最近邻[6]等分类模型来实现的。Park 等[10]提取蛋白质序列的氨基酸和氨基酸对的含量信息,并利用支持向量机算法建立的分类器,区分 OMPs 和其他蛋白质的精确度达到 94%,在已知的基于序列特征的预测方法中性能突出。此外,多种基于隐马尔科夫模型建立的预测方法[7-10]被提出来,用于预测 OMPs 跨膜区的拓扑结构。另外,有研究者利用 OMPs 的序列保守和同源信息来区分基因组内的 OMPs 和其他蛋白质[2,11-12]。

在蛋白质功能预测中,采用新的序列编码方法提高分类或预测准确度是研究热点。例如,利用氨基酸的物理化学性质以及物理学中"分组"的思想,Zhang 等[13]提出利用分组重量函数对蛋白质序列进行编码,比利用氨基酸组成信息更能有效地预测蛋白质的亚细胞位置。本研究提出利用蛋白质序列中氨基酸组成信息以及残基的理化性质,结合分组重量编码(encoding based on grouped weight, EBGW)方法来提取序列特征,并进一步利用特征选择技术来筛选特征,最后结合支持向量机算法建立分类模型,用于区分 OMPs 和其他蛋白质,并在基因组序列中预测可能的 OMPs,取得了较好的实验结果。

1 数据与方法

1.1 数据集

采用由 Gromiha 等[14]收集的蛋白质数据集,包括 OMPs(377)、TMHs(267)和 GPs(674)三大类蛋白质(括号内为序列数量),其中 OMPs 和 TMHs 两类跨膜蛋白数据从 PSORT-B 数据库中筛选而来,GPs 数据从 PDB40D_1.37 数据库中筛选而来。采用 CD-HIT 程序(http://bioinformatics.org/cd-hit/)对数据集的序列相似度进行分析后,去除序列相似度大于 40% 的冗余序列,最后得到的数据集组成如下:OMPs(208),TMHs(206),GPs(673),共 1 087 条序列,本研究称为 MCP1087 数据集。

1.2 序列分组重量编码

组成蛋白质序列的 20 种氨基酸按其理化性质可划分为 4 组:中性非极性氨基酸 G1 = {G,A,V,L,I,M,P,F,W};中性极性氨基酸 G2 = {Q,N,S,T,Y,C};酸性带负电荷的氨基酸 G3 = {D,E};碱性带正电荷的氨基酸 G4 = {H,K,R}。如果两两组合,则存在下列组合方式:{G1,G2}和{G3,G4}、{G1,G3}和{G2,G4}、{G1,G4}和{G2,G3}。设 $A = a_1, a_2, a_3, \cdots, a_n$ 是 1 条长度为 n 的蛋白质序列,可以通过定义如下 3 个映射,将序列 A 转化为 3 条 0-1 进制序列[13]:

$$\phi_1(\alpha_j) = \begin{cases} 1 & if\ \alpha_j \in \{G1, G2\} \\ 0 & if\ \alpha_j \in \{G3, G4\} \end{cases}$$

$$\phi_2(\alpha_j) = \begin{cases} 1 & if\ \alpha_j \in \{G1, G3\} \\ 0 & if\ \alpha_j \in \{G2, G4\} \end{cases} \quad (1)$$

$$\phi_3(\alpha_j) = \begin{cases} 1 & if\ \alpha_j \in \{G1, G4\} \\ 0 & if\ \alpha_j \in \{G2, G3\} \end{cases}$$

记 $H^i = \phi_i(A) = h_1^i h_2^i \cdots h_n^i (i = 1,2,3)$,称 H^1、H^2、H^3 分别为序列 A 的 1-特征序列、2-特征序列、3-特征序列。对于任意一条特征序列,定义该序列中"1"出现的次数为该序列的重量,进一步定义正规重量为该序列中"1"出现的频率。对于长度为 n 的特征序列 $H = h_1 h_{2,\cdots,} h_n$,其正规重量为 $w(n) = p/n$,其中 p 为序列中"1"出现的次数。取一个正整数 L,可将特征序列 H 划分为 L 条长度递增的子序列,记为 $H\lfloor kn/L \rfloor (k = 1,2,\cdots L)$,其中 $\lfloor \cdot \rfloor$ 表示取整运算。分别计算每条子序列的正规重量,记为 $W\lfloor kn/L \rfloor (k = 1,2,\cdots L)$,这样可以得到 1 个 L 维的向量 $W = [w\lfloor n/L \rfloor, \cdots, w\lfloor Ln/L \rfloor]$,由此,将长度为 n 的蛋白质序列编码成为一个 L 维的向量,将 H^1,H^2,H^3 等 3 条特征序列向量组合,可以得到一个 3L 维的向量,记为 $[w_1, w_2, \cdots, w_{3L}]$,称该向量为序列分组重量特征向量。此外,计算蛋白质 A 序列中 20 种氨基酸的含量,则可以得到一个 20 维的向量 $[f_1, f_2, \cdots, f_{20}]$,其中 f_i 表示第 i 种氨基酸残基在序列中的含量,称之为氨基酸组成特征向量。将分组重量特征和氨基酸组成特征进行组合,可以将一条蛋白质序列编码为一个 3L+20 维的向量:

$$X = [w_1, w_2 \cdots w_{3L}, f_1, f_2 \cdots f_{20}] \quad (2)$$

1.3 特征选择

采用 F-score[15]方法来选择特征。对于 1 个给定的 m 维训练向量 X_k,k = 1,2,…,m,若阳性和阴性的训练样本数量分别为 n+ 和 n-,则第 i 个特征的 F-score 值计算公式为:

$$F(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \quad (3)$$

这里,$\bar{x}_i$、$\bar{x}_i^{(+)}$、$\bar{x}_i^{(-)}$ 分别为第 i 个特征值的均值、正样本均值、负样本均值。$\bar{x}_i^{(+)}$ 表示第 k 个阳性样本的第 i 个特征值,$\bar{x}_i^{(-)}$ 表示第 k 个阴性样本的第 i 个特征值。分子指出了阳性数据集和阴性数据集之间的差别,而分母指出了两个数据集中每两个数据之间的差别。使用反向选择方法进行特征选择,即首先使用全部特征来训练分类模型,然后在当前特征向量中去掉 F-score 值最小的特征,重新进行训练和测试,如果分类效果提高,则舍弃该特征值,如果分类效果降低,则保留该特征值。重复这一步骤,直到预测性能无法继续提高,以此时的结果作为最终选择。

1.4 SVM 分类器

使用 LIBSVM(3.12)[16]来构建 SVM 分类器,将计算得到的蛋白质特征向量 X 作为 SVM 分类器的输入向量。使用 LIBSVM 中 python 程序包的网格搜索程序来搜索最优训练参数。

1.5 性能评价指标

在 MCP1087 数据集上进行 5 倍交叉验证测试(5-fold cross-validation tests)。使用 3 种指标来评价预测结果:敏感性(sensitivity),特异性(specificity),总精度(overall accuracy),分别定义为:

$$\text{sensitivity} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{specificity} = \frac{TN}{TN + FP} \quad (5)$$

$$\text{overall accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

这里，TP、TN、FP、FN 分别表示预测结果中真阳性、真阴性、假阳性、假阴性的数量。敏感性用于度量正确预测 OMPs 的能力，而特异性则反映了正确排除非 OMPs 的能力。MCC 相关系数定义为：

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (7)$$

MCC 的值在获得最优预测结果时为 1，在获得完全随机的预测结果时为 0。

2 结果

2.1 数据集测试结果

将数据集 MCP1087 分为 OMPs 和 GPs、OMPs 和 TMHs、OMPs 和非 OMPs（TMHs 和 GPs）3 组数据，进行 5 倍交叉验证测试，分别测试模型对每组数据中两类不同蛋白质进行区分的能力。测试结果如表 1～3 所示。在对序列进行分组重量编码的时候，测试并比较不同的 L 值下分组重量特征的预测性能，发现当各组数据的分类效果最优时，L 取值在[8,20]之间，如区分 OMPs 和非 OMPs 的时候，L 的取值为 15 时分类效果最优。训练 SVM 分类器的时候，使用不同的核函数进行测试，选择获得最佳 MCC 值的 RBF 核函数作为核函数。采用网格搜索方法，在预先设定的惩罚因子 C 和核函数 γ 值的取值范围内，试验不同的 C 值和 γ 值组合的预测性能，从中选择取得最佳预测性能的 C 值和 γ 值。

表 1 MCP1087 数据集上区分外膜蛋白和球状蛋白

序列特征	预测指标（%）			
	敏感性	特异性	总精度	马氏相关系数
分组重量编码(L=15, 45D)[a]	80.3	91.7	89.0	0.703
氨基酸组成 (20D)	82.7	93.8	91.1	0.757
分组重量编码+氨基酸组成(L=15,65D)	89.9	96.1	94.7	0.853
分组重量编码+氨基酸组成(L=15, 57D)[b]	91.9	96.9	95.7	0.901

a：括号内 L 为分组重量编码的分组值，最后的数字表示特征数量；b：进行特征选择后，获得最佳预测精度的特征数量为 57，去掉的特征为分组重量编码的第 17、18、19、32、33 个特征，以及氨基酸 Phe、Asp、Ala 的含量特征

表 2 MCP1087 数据集上区分外膜蛋白和 α 螺旋跨膜蛋白

序列特征	预测指标（%）			
	敏感性	特异性	总精度	马氏相关系数
分组重量编码(L=10, 50D)	94.7	89.3	92.0	0.841
氨基酸组成(20D)	98.6	91.7	95.2	0.905
分组重量编码+氨基酸组成(L=10, 50D)	99.0	93.7	96.4	0.929
分组重量编码+氨基酸组成(L=10, 46D)[a]	99.0	94.7	96.9	0.938

a：进行特征选择后，获得最佳预测精度的特征数量为 46，去掉的特征为分组重量编码的第 1 个特征，以及氨基酸 Gly、Ala、Pro 的含量特征

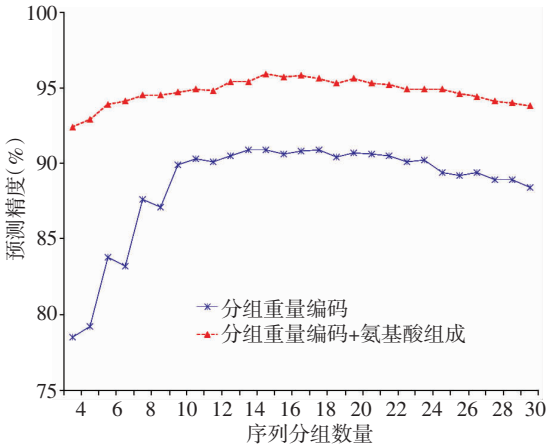
表 3 MCP1087 数据集上区分外膜蛋白和非外膜蛋白的结果

序列特征	预测指标（%）			
	敏感性	特异性	总精度	马氏相关系数
分组重量编码(L=15 42D)	87.0	91.8	90.9	0.734
氨基酸组成 n(20D)	87.5	92.7	91.7	0.754
分组重量编码+氨基酸组成(L=15, 65D)	93.3	94.9	94.6	0.837
分组重量编码+氨基酸组成(L=15, 57D)[a]	95.2	96.1	95.9	0.874

a：进行特征选择后，获得最佳预测精度的特征数量为 57，去掉的特征为分组重量编码的第 6、7、8、16、17、18、32 个特征，以及氨基酸 Ala 的含量特征

为了测试不同的 L 值下 EBGW 特征对分类精度的影响，我们绘制了 EBGW 分类精度曲线，如图 1 所示。图 1 显示使用单一的 EBGW 特征来对 OMPs 和非 OMPs 进行分类，L 的取值对分类结果的影响很大（蓝色曲线），当 L 的取值在[14,20]之间时获得最好的分类效果。而当加入氨基酸组成特征以后，由于加入的特征与 L 的取值无关，因此分类精度的变化幅度显著减小（红色曲线）。这一差异表明 EBGW 和氨基酸组成两类特征从不同的方面利用了序列中的信息，组合后可以减小其中某类特征的变化对分类性能的影响，有效地提高分类模型的鲁棒性。

模型区分 OMPs 和非 OMPs 的 ROC 曲线表明[序列编码方法为 EBGW + AA(L=15,57D)，SVM 参数 c=2,g=0.5]，当真阳性率超过 80% 时，假阳性率增加幅度很小，模型依然保持着很高的特异性。真阳性率达到 95% 左右时，假阳性率低于 20%（图 2）。ROC 曲线下面积（AUC）达到了 0.9 650，表明模型区分这两类蛋白质的准确性很高。



蓝色曲线表示 L 取值在[4,30]之间时的 EBGW 特征区分 OMPs 和非 OMPs 的精度变化情况；红色曲线表示 L 取值在[4,30]之间时，EBGW 特征和氨基酸组成特征的组合以后区分 OMPs 和非 OMPs 的精度变化情况

图 1 不同的 L 值计算的 EBGW 特征区分 OMPs 和非 OMPs 的精度变化曲线

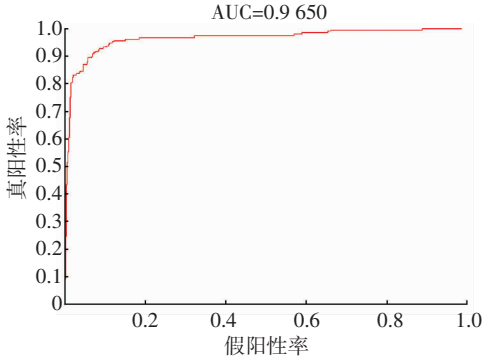


图 2 预测模型区分 OMPs 和非 OMPs 的 ROC 曲线

2.2 基因组内预测 OMPs 的结果

使用 208 条 OMPs 数据作为正训练样本，在 SCOP 数据库（1.75）中随机挑选非 OMPs 折叠类型的 450 条其他蛋白质作为负训练样本，采用 EBGW + 氨基酸组成对蛋白质序列编码，采用 F-score 选择特征子集，然后训练 SVM 分类器。在 NCBI

中下载 13 个物种基因组编码的蛋白质序列(ftp://ftp.ncbi.nih.gov/genomes/),使用最优训练参数下的模型在其中预测 OMPs。预测时,每 1 条蛋白质特征向量都以训练数据作为背景进行标准化,预测结果如表 4 所示。

在革兰阴性菌基因组内,据估计有 1.5% ~3% 的蛋白质为 OMPs[1]。我们在 4 个革兰阴性菌基因组中预测的 OMPs 的比例与这一估计相符,这其中包含了 52 个来自训练样本的蛋白,以及 212 个新预测到的 OMPs。在具备外膜结构的聚球蓝藻基因组内预测到 21 个 OMPs,表明预测模型的敏感性比较高。在不存在外膜蛋白质的 2 个古细菌和 4 个革兰阳性菌基因组内,一共预测到 0.24%(37/15 616)的 OMPs,这一比例远低于在革兰阴性菌中的预测结果,表明预测模型具有很好的特异性。此外,在真核物种中的预测结果显示该预测模型也能够发现其中的外膜蛋白,如在酵母中预测到了 6 个阳性结果,其中 4 个为已知的线粒体外膜蛋白,其中包括 2 个 VDAC 异构体蛋白。此外,在人类基因组中预测到 31 个 OMPs,其中 13 个为已知的线粒体外膜蛋白,18 个为假阳性结果,表明在与细菌差异较大的人类基因组内,预测模型的特异性降低,但是仍然保持了较好的敏感性。

表 4 在 13 个物种基因组蛋白质序列中预测 OMPs 的结果

物种类别	物种名称	蛋白质数量	预测的 OMPs 数量[例(%)]
古细菌	嗜热古菌	1700	3(0.18)
	甲烷暖球菌	1714	1(0.05)
革兰阳性菌	金黄色葡萄球菌	2650	11(0.42)
	枯草杆菌	4176	9(0.22)
	白喉棒状杆菌	2361	9(0.38)
	干酪乳杆菌	3015	4(0.13)
革兰阴性菌	大肠杆菌	4146	92(2.22)
	脑膜炎奈瑟菌	2020	54(2.67)
	根瘤土壤杆菌	5355	86(1.61)
	汉赛巴尔通体	1488	32(2.15)
蓝藻类	聚球藻	2892	21(0.73)
真核生物	酿酒酵母	5904	6(0.1)
	人	32799	31(0.09)

2.3 Web 预测程序

我们设计了算法和程序对预测模型进行了实现,开发了相应的 web 预测程序,访问地址为:http://bioinfo.tmmu.edu.cn/EBGW_OMP。用户在网页上输入蛋白质序列即可进行预测,预测程序将根据预测模型输出的概率判断该蛋白质是否为 OMP。

3 讨论

我们利用 EBGW 和氨基酸组成统计方法提取 OMPs 的序列特征,利用特征选择技术筛选更有效区分 OMPs 和其他蛋白质的特征子集,并结合性能得到广泛验证的支持向量机算法建立分类模型,利用不同的数据来验证模型正确预测 OMPs 的性能,测试结果证明这种预测方法具有很好的敏感性和特异性。

Bagos 等[10]提出的分类模型也完全基于序列特征构建,在相同的 MCP1087 数据集上,区分 OMPs 和 GPs、OMPs 和 TMHs、OMPs 和非 OMPs 的精度分别为 94.4%、95.9% 和 93.9%,我们的方法在同样的测试中精度分别提高了 1.3%、1.0% 和 2.0%。Bagos 等[10]采用的 FDR(fisher discriminant ratio)来选择特征子集,与 F-score 方法相比,对于不同特征的贡献大小的计算结果极为接近,说明预测精度的差异主要来源于序列编码策略的不同。EBGW 编码方法能够提取蛋白质序列不同区域上不同性质的氨基酸出现频率的差异信息。OMPs 中跨膜区为 β 折叠链,TMHs 的跨膜区为 α 螺旋,而二者跨膜区的长度则很接近,在合理的分组数量下,EBGW 能够获取跨膜区序列特征上的差异,而 Bagos 等[10]使用的氨基酸对特征则不能体现这种差异。此外,将 EBGW 特征与氨基酸组成特征组合以后,预测精度比单独使用一类特征显著提高,说明二者相互补充,从不同方面揭示了序列中的信息。

OMPdb 数据库[17]中收录了 5 265 条 OMPs 数据(序列间相似度低于 30%),去掉 146 条包含在 208 条训练样本中的序列,在余下的 5 119 个蛋白质中,我们的方法预测到 4 566 个 OMPs,真阳性率为 89.2%,Bagos 等[10]的方法在其中预测到 4 472 个 OMPs,真阳性率为 87.4%,敏感性低于我们的方法,这与 MCP1087 上的测试结果一致。考虑到训练样本的不同将导致预测性能的差异,我们没有在 13 个基因组测试中同样测试和比较 Bagos 等[10]的方法。

本研究表明,完全从序列中提取特征,通过改进序列编码方法,在目前能够利用的分类算法的限制下,分类性能的提升较为有限。如果要更为准确地在基因组中预测 OMPs,需要进一步利用进化上和结构上的信息。完全基于序列特征的分类模型,其优势是计算速度更快,适合于基因组内大规模筛选。

参考文献:

[1] Wimley W C. The versatile beta-barrel membrane protein[J]. Curr Opin Struct Biol, 2003, 13(4): 404-411.

[2] Punta M, Forrest L R, Bigelow H, *et al*. Membrane protein prediction methods[J]. Methods, 2007, 41(4): 460-474.

[3] Gromiha M M, Ahmad S, Suwa M. Neural network-based prediction of transmembrane beta-strand segments in outer membrane proteins[J]. J Comput Chem, 2004, 25(5): 762-767.

[4] Jacoboni I, Martelli P L, Fariselli P, *et al*. Prediction of the transmembrane regions of beta-barrel membrane proteins with a neural network-based predictor[J]. Protein Sci, 2001, 10(4): 779-787.

[5] Park K J, Gromiha M M, Horton P, *et al*. Discrimination of outer membrane proteins using support vector machines[J]. Bioinformatics, 2005, 21(23): 4223-4229.

[6] Garrow A G, Agnew A, Westhead D R. TMB-Hunt: a web server to screen sequence sets for transmembrane beta-barrel proteins[J]. Nucleic Acids Res, 2005, 33(Web Server issue): W188-W192.

[7] Bigelow H R, Petrey D S, Liu J, *et al*. Predicting transmembrane beta-barrels in proteomes[J]. Nucleic Acids Res, 2004, 32(8): 2566-2577.

[8] Liu Q, Zhu Y S, Wang B H, *et al*. A HMM-based method to predict the transmembrane regions of beta-barrel membrane proteins[J]. Comput Biol Chem, 2003, 27(1): 69-76.

[9] Bagos P G, Liakopoulos T D, Spyropoulos I C, *et al*. PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins[J]. Nucleic Acids Res, 2004, 32(Web Server issue): W400-W404.

[10] Bagos P G, Liakopoulos T D, Hamodrakas S J. Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method[J]. BMC Bioinformatics, 2005, 6: 7.

[11] Garrow A G, Westhead D R. A consensus algorithm to screen genomes for novel families of transmembrane beta barrel proteins[J]. Proteins, 2007, 69(1): 8-18.

[12] Berven F S, Flikka K, Jensen H B, *et al*. BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria[J]. Nucleic Acids Res, 2004, 32(Web Server issue): W394-W399.

[13] Zhang Z H, Wang Z H, Zhang Z R, *et al*. A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine[J]. FEBS Lett, 2006, 580(26): 6169-6174.

[14] Gromiha M M, Suwa M. A simple statistical method for discriminating outer membrane proteins with better accuracy[J]. Bioinformatics, 2005, 21(7): 961-968.

[15] Chen Y W, Lin C J. Combining SVMs with Various Feature Selection Strategies[J]. Studies in Fuzziness and Soft Computing, 2006, 207: 315-324.

[16] Chang C C, Lin C J. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3): 27.

[17] Tsirigos K D, Bagos P G, Hamodrakas S J. OMPdb: a database of β-barrel outer membrane proteins from Gram-negative bacteria[J]. Nucleic Acids Res, 2011, 39(Database issue): D324-D331.

(**收稿**:2012-12-27;**修回**:2013-01-24)

(编辑 吴培红)

(上接1361页)

上存在交叉点,两者有共同的细胞起源,提示两种疾病的并发存在细胞学及分子生物学[1]上的依据;本例患者在移植39个月后发现两种恶性克隆细胞,但其发生的先后顺序及是否源于同一种病因还无法确证。目前不少学者认为,治疗相关性白血病的发生与移植期间烷化剂和拓扑异构酶抑制剂的使用存在一定的相关性,机制考虑为:该类药物破坏细胞DNA,导致基因组不稳定,继而激活致癌基因Ras,使抑癌基因P53发生突变,导致细胞增殖分化失控。该患者在MM的前期化疗和移植预处理方案中使用过多种烷化剂和拓扑异构酶抑制剂(环磷酰胺、马法兰、瘤可宁等),且累计剂量偏大,均可能成为诱导并发ALL的危险因素。如何预防MM移植后并发急性白血病,国内外尚无统一定论,我们认为,探索减少烷化剂和拓扑异构酶抑制剂的累计剂量,可能起到积极预防作用,为不影响疗效,化疗过程中可辅以硼替佐米等靶向治疗。今后临床医师需收集该类病例,探讨预处理方案中各药物类型及剂量的选择,平衡药物作用及不良反应的关系,提高自体造血干细胞移植治疗多发性骨髓瘤的有效性和安全性。

据国内外统计,约有5%的多发性骨髓瘤可转变为骨髓增生异常综合征(myelodysplastic syndromes,MDS)或急性白血病(acute leukemia,AL),而AL中又以急性非淋巴细胞白血病(acute nonlymphocytic leukemia,ANLL)最为多见。多发性骨髓瘤化疗后发生AL的高峰通常在用药2年左右开始增高,5~10年达到高峰。本例患者自体造血干细胞移植后39个月复发合并急性B淋巴细胞白血病(B-ALL),且发病时间有所提前。

国内曹静等[2]报道1例IgD型MM经CD、MP、VAD方案化疗后达CR,5年后转为ALL(L2),经DVD、Hyper-CVAD、HD-MTX/Ara-c诱导后达完全缓解,6个月后行自体造血干细胞移植(Allo-HSCT),抗排斥反应治疗中。李志婷等[3]报道1例IgA型MM经COMP和MP方案化疗后达CR,5年后转为ALL(L_1),VP方案诱导2个疗程后死于呼吸衰竭。张中华等[4]报道1例IgD型MM经M2方案化疗后达CR,13个月后转为ALL(L_2)。而经过自体造血干细胞移植后并发急性淋巴细胞白血病的病例国内鲜见报道。TsukadaY等[5]曾报道1例MM经自体外周血干细胞移植及马法兰化疗后达CR,5年后转为急性B淋巴细胞白血病(B-ALL)。本病例多发性骨髓瘤自体造血干细胞移植后复发合并急性淋巴细胞白血病L_2型。由于患者经济受限,前期针对MM治疗时无法予以万珂进行联合化疗,仅以VTP、CHOD方案化疗及对症治疗,多次复查骨髓象提示未缓解。在临床工作中,面临此类患者时可考虑行MM和ALL耐药基因的检测,条件许可时,还可进行肿瘤细胞药物敏感试验,尝试难治性MM和ALL治疗方案的组合,寻找治疗此类疾病的有效方法。

参考文献:

[1] Mailankody S, Pfeiffer R M, Kristinsson S Y, *et al*. Risk of acute myeloid leukemia and myelodysplastic syndromes after multiple myeloma and its precursor disease(MGUS)[J]. Blood, 2011, 118(15): 4086-4092.

[2] 曹静, 吴小津, 唐晓文, 等. 多发性骨髓瘤继发急性淋巴细胞性白血病1例并文献复习[J]. 现代医药卫生, 2007, 23(10): 1440-1442.

[3] 李志婷, 刘连寅. 多发性骨髓瘤继发急性淋巴细胞白血病1例[J]. 临床荟萃, 2005, 20(19): 1123.

[4] 张中华, 郝 军. IgD型多发性骨髓瘤并发急性淋巴细胞性白血病1例[J]. 肿瘤学杂志, 2005, 11(1): 74.

[5] Tsukada Y, Hattori Y, Nakajima H, *et al*. B-cell acute lymphoblastic leukemia developed 5 years after autologous stem cell transplantation for multiple myeloma[J]. Rinsho Ketsueki, 2012, 53(2): 219-223.

(**收稿**:2013-04-05;**修回**:2013-05-14)

(编辑 吴培红)