生 物 工 程 学 报
journals.im.ac.cn
cjb@im.ac.cn

*Chin J Biotech* 2008, August 25; 24(8): 1439-1445
Chinese Journal of Biotechnology ISSN 1000-3061


研究报告

# 基于修正的伪氨基酸组成预测氧化还原酶辅酶类型

张光亚, 李红春, 方柏山

华侨大学工业生物技术研究所, 泉州 362021

**摘　要:** 从序列出发快速确定氧化还原酶的辅酶依赖类型对于了解其结构和功能、催化机制及构建辅酶再生体系具有重要指导作用。对 Chou 提出的伪氨基酸组成方法进行了修正并用于提取氧化还原酶序列特征值, 采用 *k*-近邻算法预测其辅酶依赖类型。当 λ=48, *w*=0.1 时, 10 倍交叉验证结果表明: 其 ROC 曲线下面积为 0.9536, 预测精度达 92.0%, 比最优条件下伪氨基酸组成预测精度提高了 3.5%; 与其他 7 种常见特征值提取方法相比, 修正的伪氨基酸组成表现最好。结果表明从序列出发预测氧化还原酶辅酶依赖类型是可行的, 且修正的伪氨基酸组成可望成为一种新的有效提取蛋白质序列特征值方法。

**关键词:** 氧化还原酶, 辅酶依赖类型, 修正的伪氨基酸组成, *k*-近邻, ROC 曲线下面积

# Predicting the Cofactors of Oxidoreductases by the Modified Pseudo-amino Acid Composition

Guangya Zhang, Hongchun Li, and Baishan Fang

*Institute of Industrial Biotechnology*, *Huaqiao University*, *Quanzhou* 362021, *China*

**Abstract:** Types of cofactor independency for newly found oxidoreductases sequences are usually determined by experimental analysis. These experimental methods are both time-consuming and costly. With the explosion of oxidoreductases sequences entering into the databanks, it is highly desirable to explore the feasibility of selectively classifying newly found oxidoreductases into their respective cofactor independency classes by means of an automated method. In this study, we proposed a modified Chou's pseudo-amino acid composition method to extract features from sequences and the *k*-nearest neighbor was used as the classifier, and the results were very encouraging. When $\lambda$=48, $w$=0.1, the areas under the ROC curve of *k*-nearest neighbor in 10-fold cross-validation was 0.9536; and the success rate was 92.0%, which was 3.5% higher than that of pseudo-amino acid composition. It was also better than all the other 7 feature extraction methods. Our results showed that predicting the cofactors of oxidoreductases was feasible and the modified pseudo-amino acid composition method may be a useful method for extracting features from protein sequences.

**Keywords**: oxidoreductases, cofactor independency type, modified pseudo-amino acid composition, *k*-nearest neighbor, areas under the ROC curve

在国际生物化学联合会划分的 6 大类酶中, 约　35%为氧化还原酶, 由于氧化还原酶在催化制备手

**Received**: December 10, 2007; **Accepted:** February 27, 2008
**Supported by**: the Natural Science Foundation of Fujian Province (No. 2007J0360) and the Research Fund for the Doctoral Program of Higher Education (No. 20070385001).
**Corresponding author:** Baishan Fang. Tel: +86-595-22691560; E-mail: fangbs@hqu.edu.cn
福建省自然科学基金项目(No. 2007J0360), 高等学校博士学科点专项科研基金项目(No. 20070385001)资助。

性醇、羟基酸、氨基酸方面显示出极大的优势, 因此在制药、食品、精细化工、农药等领域具有重要的用途; 且因其能够选择性地催化有机物中的相应基团, 往往能够一步催化完成非生物催化剂需经多步催化才能完成的复杂化学反应, 故其应用越来越受到人们的重视[1,2]。而在催化过程中, 几乎所有的氧化还原酶均需要辅酶的参与, 这些辅酶包括NAD+, NADP+, FAD, FMN以及它们的还原态等。

近年来, 随着测序技术的迅猛发展, 数据库中蛋白质序列飞速增长, 1986 年SWISS-PROT仅收录了 3939 条序列[3], 而目前已经收录了 270778 序列(release 53.1, 2007 年 6 月 12 日), 在短短的 20 年中增加了 67.7 倍。对于一个新获得的蛋白质序列, 研究者首先要考虑的是它是不是一个酶? 如果是, 是什么酶? 它催化的机理是什么? 需不需要辅酶? 需要什么辅酶? 而后者对于氧化还原酶尤为重要, 因为这涉及到对酶结构和功能的理解以及催化机制的探讨, 且对构建什么样的辅酶再生体系具有指导作用。尽管上述问题都可通过不同的生化实验进行证实, 但完全基于实验的方法既费时费力, 而且由于辅酶非常昂贵, 使该过程花费较大。因此, 研究者需要一种快速、自动化的方法迅速从酶的序列出发, 了解其依赖什么类型的辅酶, 进而指导其实验。有研究者[4]从短链氧化还原酶(SCOR)3D结构出发, 通过计算静电引力能量占辅酶-酶分子相互作用总能量的比例, 并以此对其NAD/NADP两种辅酶的依赖情况进行预测, 预测精度最高可达 90%。众所周知, 获取酶分子 3D结构远比获取其序列困难, 因此, 在实际应用中受到限制; 最近, Yvonne Kallberg等[5]利用隐马式模型通过搜索氧化还原酶序列中的Rossmann折叠的特征性序列GxGxxG(其中x代表任意氨基酸), 对其辅酶FAD、NAD和NADP进行预测, 取得了较好的识别效果, 但他们同时也指出, 并非所有的氧化还原酶都具有此特征性序列, 在SCOR中有 26%的酶具有此序列, 为拥有此结构最多的一种氧化还原酶类型。可见, 该方法在应用中依然存在局限。此外, 借助序列比对(序列间的相似性)预测辅酶虽具有一定的指导和借鉴作用, 但有些氧化还原酶可能找不到对应的同源物, 而且参照比对的结果有时不太准确。如: 文献报道[6]: 野生型Lactaldehyde dehydrogenase (ALD, 丙酮醛脱氢酶)只利用NAD作为辅酶, 而仅将其活性部位的F改变为T(F180T)后, 就可以利用NADP作为辅酶, 同时对NAD的亲和能力也显著增强。

本研究获得了 2462 条氧化还原酶序列, 并根据其辅酶依赖类型分为 5 类, 完全基于其序列特征, 提出了修正的伪氨基酸组成方法并用于提取序列特征值, 结合 *k*-近邻算法对其辅酶依赖类型进行了预测, 取得了令人满意的结果, 现报道如下。

## 1 材料和方法

### 1.1 数据来源

样本中氧化还原酶来源于 ENZYME 数据库[7], 其序列来源于 Swiss-Prot[8], 所有序列根据其辅因子不同分为 5 类: 1) FAD, 指该酶仅使用 FAD 或其还原态作为辅因子(共计 969 条); 2) FMN, 指该酶仅使用 FMN 或其还原态作为辅因子(共计 453 条); 3) heme, 指该酶仅使用 heme 作为辅因子(共计 220 条)。4). NAD, 指该酶仅使用 NAD 或其还原态作为辅因子(共计 1411 条); 5) NADP, 指该酶仅使用 NADP 或其还原态作为辅因子(共计 1616 条); 剔除了其中长度小于 100 个氨基酸残基的酶蛋白和部分片断酶蛋白; 并使用 BLASTCLUST 程序[9]剔除了样本所有相似性大于 30%的序列, 使任意两条序列之间的全同率(Sequence identity)均小于 30%, 以提高识别效果的可靠性, 该方法已得到较广泛的应用[10,11]。最后五类氧化还原酶的样本量分别为 378、151、166、715 和 1052, 共计 2462 条。样本中序列长度的分布情况

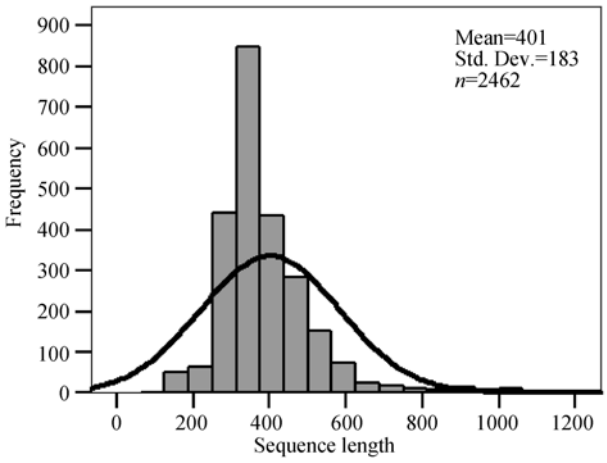


**图 1 样本中氧化还原酶长度的分布**

**Fig. 1 Length distribution in the oxidoreductases sample**

如图 1 所示, 可见大多数氧化还原酶长度都在300~500 个氨基酸之间。

### 1.2 修正的伪氨基酸组成

伪氨基酸组成(psaa)最早由Chou等[12]提出, 对于一个长度为L的蛋白质序列$R_1R_2R_3R_4R_5R_6R_7\cdots R_L$, 其序列顺序的影响可通过一系列的序列顺序相关因子(Order-correlation factor)进行表述:

$$\begin{cases}\theta_1=\dfrac{1}{L-1}\sum\limits_{i=1}^{L-1}\Theta(R_i,R_{i+1})\\ \theta_2=\dfrac{1}{L-2}\sum\limits_{i=1}^{L-2}\Theta(R_i,R_{i+2})\\ \theta_3=\dfrac{1}{L-3}\sum\limits_{i=1}^{L-3}\Theta(R_i,R_{i+3})\\ \cdots\\ \theta_\lambda=\dfrac{1}{L-\lambda}\sum\limits_{i=1}^{L-\lambda}\Theta(R_i,R_{i+\lambda})\end{cases}\tag{1}$$

(1)式中, $\theta_1$表示第一等级(First-tier)相关因子, 表示蛋白质序列中相邻氨基酸相关性, $\theta_2$表示第二等级相关因子, 表示蛋白质序列中每间隔一个氨基酸之间的相关性, 其他依此类推; Θ表示相关性函数, 其定义为:

$$\Theta(R_i,R_j)=\frac{1}{3}\begin{Bmatrix}\left[H_1(R_j)-H_1(R_i)\right]^2\\ +\left[H_2(R_j)-H_2(R_i)\right]^2\\ +\left[M(R_j)-M(R_i)\right]^2\end{Bmatrix}\tag{2}$$

式中, $H_1(R_i)$, $H_2(R_i)$和$M(R_i)$分别表示氨基酸$R_j$的疏水性、亲水性和侧链分子量, $H_1(R_j)$, $H_2(R_j)$和$M(R_j)$分别表示氨基酸$R_j$的疏水性、亲水性和侧链分子量。将氨基酸的疏水性、亲水性和侧链分子量带入公式 2 之前, 将它们按照公式 3 进行标准化转换:

$$\begin{cases}H_1(i)=\dfrac{H_1^0(i)-\sum\limits_{i=1}^{20}\dfrac{H_1^0(i)}{20}}{\sqrt{\dfrac{\sum\limits_{i=1}^{20}\left[H_1^0(i)-\sum\limits_{i=1}^{20}\dfrac{H_1^0(i)}{20}\right]^2}{20}}}\\ H_2(i)=\dfrac{H_2^0(i)-\sum\limits_{i=1}^{20}\dfrac{H_2^0(i)}{20}}{\sqrt{\dfrac{\sum\limits_{i=1}^{20}\left[H_2^0(i)-\sum\limits_{i=1}^{20}\dfrac{H_2^0(i)}{20}\right]^2}{20}}}\\ M(i)=\dfrac{M^0(i)-\sum\limits_{i=1}^{20}\dfrac{M^0(i)}{20}}{\sqrt{\dfrac{\sum\limits_{i=1}^{20}\left[M^0(i)-\sum\limits_{i=1}^{20}\dfrac{M^0(i)}{20}\right]^2}{20}}}\end{cases}\tag{3}$$

式中, $H_1^0(i)$表示第 $i$ 个氨基酸原始疏水性值, $H_2^0(i)$表示第 $i$ 个氨基酸原始亲水性值, $M^0(i)$表示第 $i$ 个氨基酸原始侧链分子量。然后, 将公式(1)计算的结果与 20 种氨基酸组成进行离散化处理。这样, 一个蛋白质 $X$ 序列就可通过 20+$\lambda$维的数组表示(公式 4), (4)式中

$$X=\begin{bmatrix}x_1\\ \vdots\\ x_{20}\\ x_{20+1}\\ \vdots\\ x_{20+\lambda}\end{bmatrix}\tag{4}$$

公式(5)中, $f_i$表示蛋白质$X$经过归一化处理后的20 种氨基酸组成, $\theta_j$表示第$j$等级序列相关因子(由公式 1 计算), $w$为权重因子。在计算过程中, $\lambda$和$w$需要进行选择, 以达到最佳识别效果。

$$x_u=\begin{cases}\dfrac{f_u}{\sum\limits_{i=1}^{20}f_i+w\sum\limits_{j=1}^{\lambda}\theta_j},(1\leqslant u\leqslant 20)\\ \dfrac{w\theta_u-20}{\sum\limits_{i=1}^{20}f_i+w\sum\limits_{j=1}^{\lambda}\theta_j},(20+1\leqslant u\leqslant 20+\lambda)\end{cases}\tag{5}$$

我们在计算过程中对公式 2 和 3 分别进行了修正, 如公式 6 和 7 所示。

$$\Theta(R_i,R_j)=\frac{1}{3}\begin{Bmatrix}\left[Z_1(R_j)-Z_1(R_i)\right]^2+\left[Z_2(R_j)-Z_2(R_i)\right]^2+\\ \left[Z_3(R_j)-Z_3(R_i)\right]^2\end{Bmatrix}\tag{6}$$

公式 6 中, $Z_1(R_i)$, $Z_2(R_i)$和$Z_3(R_i)$分别表示氨基酸$R_j$$Z$标度值的前 3 个主成分; 公式 7 中, $Z_1^0(i)$表示第$i$个氨基酸原始$Z$标度值第 1 个主成分值, $Z_2^0(i)$表示第$i$个氨基酸原始$Z$标度值第 2 个主成分值, $Z_3^0(i)$表示第$i$个氨基酸原始$Z$标度值第 3 个主成分值。

$Z$标度($Z$-scales)是Hellberg等[13]对氨基酸 29 个物理化学性质进行主成分分析, 得到了 3 个显著的主成分, 并将相应主成分得分矢量作为新的氨基酸描

述子, 称之为主性质(Principal property), 即*Z*标度, 该方法在构建多肽定量构效关系模型等领域得到了广泛应用[14]。

$$\begin{cases} Z_1(i)=\dfrac{Z_1^0(i)-\sum_{i=1}^{20}\dfrac{Z_1^0(i)}{20}}{\sqrt{\dfrac{\sum_{i=1}^{20}\left[Z_1^0(i)-\sum_{i=1}^{20}\dfrac{Z_1^0(i)}{20}\right]^2}{20}}} \\ Z_2(i)=\dfrac{Z_2^0(i)-\sum_{i=1}^{20}\dfrac{Z_2^0(i)}{20}}{\sqrt{\dfrac{\sum_{i=1}^{20}\left[Z_2^0(i)-\sum_{i=1}^{20}\dfrac{Z_2^0(i)}{20}\right]^2}{20}}} \\ Z_3(i)=\dfrac{Z_3^0(i)-\sum_{i=1}^{20}\dfrac{Z_3^0(i)}{20}}{\sqrt{\dfrac{\sum_{i=1}^{20}\left[Z_3^0(i)-\sum_{i=1}^{20}\dfrac{Z_3^0(i)}{20}\right]^2}{20}}} \end{cases} \quad (7)$$

### 1.3 有效性检验

模型的稳定性及泛化能力采用了较为客观和严格的交叉验证(Cross-validation)方法, 根据相关文献[15]采用了 10 倍交叉验证(10-fold cross-validation, 10-CV), 具体做法是: 将训练数据随机分为 10 组, 每次留出 1 组作为测试数据, 另 9 组作为训练数据, 这样循环 10 次, 使得每组数据都能作为测试数据进行预测。

### 1.4 识别效果评估

模型最终表现通过以下 2 个参数进行描述, 预测正确率(Success rate), 指正确预测的辅酶依赖类型占相应总数的百分比, 此外, 采用了受试者操作特性曲线下面积(Area under the receiver operation characteristic curve, AUC)作为衡量标准[16]。由于ROC能兼顾灵敏度和特异性要求以综合评价分类器的识别性能, ROC曲线下面积作为量化指标可以直观有效的比较不同分类器的性能优劣。曲线越凸说明判别模型诊断价值越高, 并可通过计算ROC曲线下面积 ($0.5 \leqslant AUC \leqslant 1$)这一综合统计量作定量分析, 任何一个随机猜测的模型其AUC值为 0.5; 一个完美的分类器其AUC值为 1, 一般而言, 一个分类器的AUC大于 0.9, 则被认为是优秀的分类器。

文中实现*k*-近邻算法的软件来自于Weka (Waikato environment for knowledge analysis), 该程序包是基于JAVA虚拟机开发的[17]。*k*-近邻法是一种简单有效的分类方法, 其基本思路是: 在给定新样本后, 考虑在训练样本集中与该新样本距离最近(最相似)的*K*个样本, 根据这*K*个样本所属的类别判定新样本所属的类别, 其详细内容可参见文献[18], 文中*K*=1, 采用了Euclidean Distance(欧氏距离)来衡量。使用的PC为DELL precision$^{TM}$490 工作站。

## 2 结果与分析

### 2.1 基于伪氨基酸组成的辅酶依赖类型的预测

如上所述, $\lambda$和 $w$ 是伪氨基酸组成中 2 个重要参数, 本试验考察了这 2 个参数对识别精度的影响, 结果如图 2 所示。从图中可以看出, 在 10 倍交叉验证过程中, $w$ 一定的条件下, 随着$\lambda$的增大, 识别精度也随之而增大, 当$\lambda$=48 时, 均达到最大值, 随后表现出精度的略微下降; 而当$\lambda$一定时, 随着 $w$ 的增大, 识别精度表现出现增后减的趋势, 一般 $w$ 取较小值时, 其识别精度相对较高。当 w=0.05, $\lambda$=48 时, 其识别精度最高, 可达 88.5%。对 5 类氧化还原酶辅酶依赖类型正确预测的数量分别为 318, 149, 142, 632 和 937 个, 正确率分别为 84.1%, 98.7%, 85.5%, 88.4% 和 89.1%。此时, 输入的变量数目为 20+48=68 个。

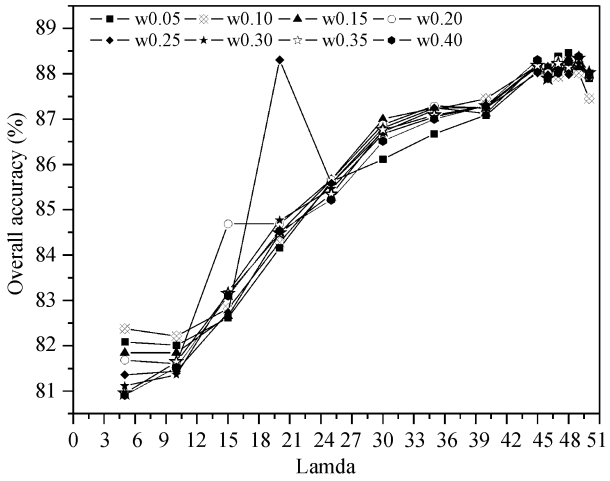


**图 2 伪氨基酸组成运行参数的选择**

**Fig. 2 Selection of running parameters of psaa**

### 2.2 基于修正的伪氨基酸组成的辅酶依赖类型的预测

在我们提出的修正的伪氨基酸组成(Mpsaa)中, $\lambda$和 $w$ 同样是 2 个重要参数, 通过 10 倍交叉验证

考察了这 2 个参数对识别精度的影响, 结果如图 3 所示。从图中可以看出, 在 $w$ 一定的条件下, 随着$\lambda$的增大, 识别精度也随之而增大, 当$\lambda$=48 时, 几乎均达到最大值, 随后表现出精度的略微下降; 而当$\lambda$一定时, 随着 $w$ 的增大, 识别精度表现出现增后减的趋势, 一般 $w$ 取较小值时, 其识别精度相对较高。其变化趋势与伪氨基酸组成类似。而当 w=0.1, $\lambda$=48 时, 其识别精度最高, 可达 92.0%。对 5 类氧化还原酶辅酶依赖类型正确预测的数量分别为 345, 151, 149, 657 和 963 个, 正确率分别为 91.3%, 100.0%, 89.8%, 91.9%和 91.5%。可见, 采用修正的伪氨基酸组成对 5 种氧化还原酶辅酶依赖类型预测精度均有不同程度的提高, 提高幅度最大为 7.2%, 整体识别精度提高了 3.5%。此时, 变量数目同样为 68。

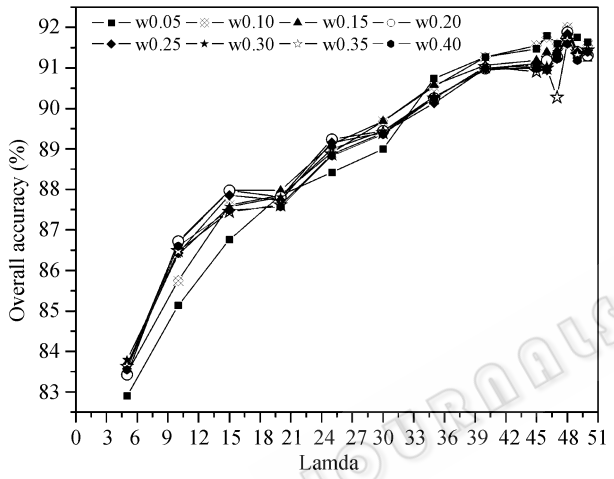


**图 3 修正的伪氨基酸组成运行参数的选择**

**Fig. 3 Selection of running parameters of Mpsaa**

为了进一步比较 2 种方法, 在相同参数条件下, 对psaa和Mpsaa的预测精度进行了比较, 结果如图 4 所示。可以发现, 随着$\lambda$和$w$的增大, 二者变化趋势相同, 但在相同条件下, Mpsaa识别精度均高于psaa约 3.3%左右。另外, 从图中可以看出, 无论在Mpsaa或psaa中, 相比于参数$w$, $\lambda$的变化对识别精度的影响更大, 这与Chou等人[19]报道结果吻合。

### 2.3 与其他序列特征值提取方法的比较

为了考察修正的伪氨基酸组成的有效性, 利用同样的数据以及$k$-近邻算法(运行参数相同), 采用了其他的序列特征值提取方法, 这些特征值提取方法包括: 氨基酸组成、二肽组成、标准化的Moreau-Broto自相关指数(Normalized Moreau-Broto autocorrelation descriptors)[20], Moran自相关指数(Moran autocorrelation descriptoers)[21], Geary自相关指数(Geary autocorrelation descriptoers)[22], 组成、转变及分布[23]以及序列顺序耦合数和部分序列顺序[24]。经 10 倍交叉验证, 其结果见表 1。

由表 1 可知, 在 9 种不同特征值提取方法中, 我们提出的修正的伪氨基酸组成识别精度最高, 部分序列顺序次之, 伪氨基酸组成再次之, 而氨基酸组成和二肽组成识别精度较低, 均在 80%以下。对第 1 类型氧化还原酶而言, Mpsaa 比二肽组成预测精度高出了 35.2%; 对第 2 类型氧化还原酶而言, Mpsaa 比二肽组成预测精度提高了 15.9%; 对第 3 类型氧化还原酶而言, Mpsaa 比二肽组成预测精度提高了 38.6%; 对第 4 类型氧化还原酶而言, 其预测精度比二肽组成提高了 15.3%; 对第 5 类型氧化还原酶而言, 其预测精度比氨基酸组成提高了 10.2%。除了对第 5 类型氧化还原酶预测精度略低于(0.3%)Geary 自

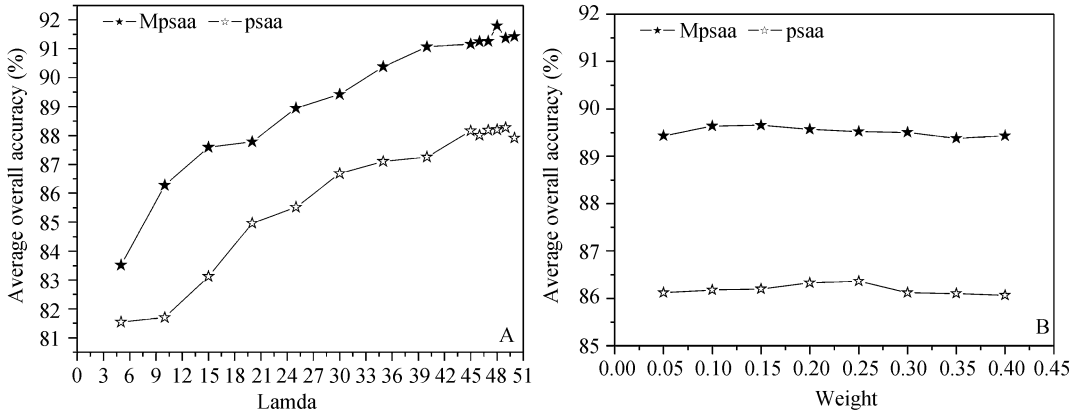


**图 4 相同运行参数下两种方法预测精度的比较**

**Fig. 4 Comparison of the two methods based on the same running parameters**

表 1 修正的伪氨基酸组成与其他特征值提取方法的比较

**Table 1 Comparison of modified pseudo-amino acid composition with other feature extraction methods**

| Feature extraction methods | Cofactors | | | | | Accuracy (%) |
|---|---|---|---|---|---|---|
| | FAD | FMN | Heme | NAD | NADP | |
| Modified pseudo-amino acid composition (1) | 91.3 | 100.0 | 89.8 | 91.9 | 91.5 | 92.0 |
| Sequence-order-coupling number & quasi-sequence-order (2) | 89.4 | 99.3 | 89.2 | 87.3 | 90.0 | 89.6 |
| Pseudo-amino acid composition (3) | 84.1 | 98.7 | 85.5 | 88.4 | 89.1 | 88.5 |
| Normalized Moreau-Broto autocorrelation descriptors (4) | 75.4 | 98.7 | 71.1 | 85.0 | 90.0 | 85.6 |
| Composition, transition and distribution (5) | 81.2 | 98.7 | 79.5 | 84.3 | 85.6 | 85.0 |
| Geary autocorrelation descriptors (6) | 71.2 | 96.0 | 64.5 | 82.0 | 91.8 | 84.2 |
| Moran autocorrelation descriptors (7) | 70.6 | 96.0 | 65.7 | 82.4 | 91.2 | 84.0 |
| Amino acid composition (8) | 71.2 | 90.7 | 81.9 | 77.8 | 81.3 | 79.3 |
| Dipeptide composition (9) | 56.1 | 84.1 | 51.2 | 76.6 | 83.2 | 75.0 |

相关指数以外, Mpsaa 对其他 4 类氧化还原酶预测精度均为最佳。

为了进一步比较各种特征值提取方法的优劣, 在其他条件完全相同的情况下, 计算了各自 ROC 曲线下面积的平均值(见图 5, 其中特征值提取方法的序号与表 1 相同), 可见, AUC 值反映的趋势和预测精度基本一致, 即: 预测精度较高的方法其对应的 AUC 值也较大, 而 Mpsaa 方法的 AUC 值达到 0.9536, 说明此时分类器表现非常优秀。

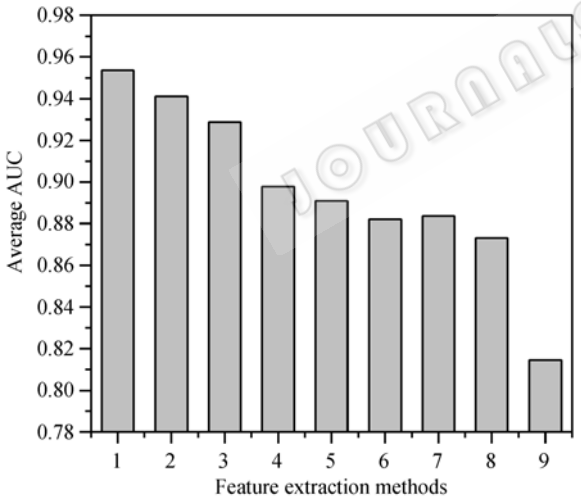


图 5 不同方法 AUC 值的比较

**Fig. 5 The AUC values of different feature extraction methods**

为了分析出现预测错误的原因, 统计了出现错误预测的情况, 结果表明: 在以氨基酸组成作为特征值时, 在所有预测错误中, $NAD^+$和 $NADP^+$型的数量占总数的 69.9%, 其中将 $NAD^+$预测为 $NADP^+$的个数为 112 个, 占总数量的 22.0%; 而将 $NADP^+$预测为 $NAD^+$的个数为 133 个, 占总数量的 26.1%, 而在以修正的伪氨基酸组成作为特征值时, 虽然预测错误数量大量减少, 但在所有预测错误中, $NAD^+$和 $NADP^+$型的数量占总数的 74.6%, 其中将 $NAD^+$预测为 $NADP^+$的个数为 45 个, 占总数量的 22.8%; 而将 $NADP^+$预测为 $NAD^+$的个数为 61 个, 占总数量的 31.0%, 类似的趋势也存在于其他特征值提取方法中。这可能是由于 $NAD^+$和 $NADP^+$在结构上非常相似, 只是相差 1 个磷酸基团, 因此导致算法在分类过程中出现错误。因此, 如何有效提高对这 2 类问题的识别能力将是后续研究的重点。

## 3 讨论

化学及生物化学的很多转化过程都涉及氧化还原反应, 开发有实际应用价值的氧化还原酶是生物技术领域一直以来的一项重要内容, 因此, 辅酶依赖型氧化还原酶在生物转化中的作用越来越受到重视, 对于新获得的氧化还原酶, 要确定其辅酶依赖类型, 首先要分离和纯化该酶, 然后通过测定不同辅酶存在时酶活性的变化来确定其辅酶[25], 显然, 这是一个费时、费力的过程, 而且由于获取氧化还原酶序列的信息变得越来越容易, 因此, 发展一种快速判断氧化还原酶辅酶依赖类型的方法就显得尤为迫切。正因为如此, 研究者对此表现出浓厚的兴趣[4,5,26], 但这些方法主要针对短链氧化还原酶(SCOR), 而且在预测过程中要么使用SCOR的晶体结构, 要么仅使用其序列中的部分特征性序列, 而且预测的辅酶类型相对较少。因此, 在实际应用过程中存在一定的缺陷。本研究拓展到所有的氧化还原酶, 而且对其常见的 5 类辅酶类型均进行了预测,

并得到了较高识别精度, 这说明从氧化还原酶序列出发, 对其辅酶依赖类型进行预测是可行的, 且可获得较高的准确率。

从蛋白质序列出发对其生物学特性进行预测是目前生物信息学研究的热点问题, 也是探讨蛋白质结构和功能关系的一种重要研究手段。如何从序列信息中有效提取特征值是目前关注的焦点之一。本文提出了一种修正的伪氨基酸组成的方法, 并应用到氧化还原酶辅酶依赖类型的预测, 取得了优于其他特征值提取方法的结果。相比于伪氨基酸组成仅考虑了氨基酸残基的亲水性、疏水性以及侧链分子量, 我们提出的修正的伪氨基酸组成则通过主成分分析对氨基酸多种理化特征常数的信息进行了浓缩, 其包含的信息量要远多于前者, 因此, 取得较好的识别效果也就顺理成章了。本研究提出修正的伪氨基酸组成可望成为一种提取蛋白质特征值的新方法, 在预测蛋白质其他特性的生物信息学研究中得到应用。

## REFERENCES


[1] Hummel W. Large-scale applications of NAD(P)-dependent oxidoreductases: recent developments. *Trends Biotechnol*, 1999, **17**(12): 487–492.

[2] Liu WF, Wang P. Cofactor regeneration for sustainable enzymatic biosynthesis. *Biotech Adv*, 2007, **25**(4): 369–384.

[3] Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res*, 2000, **25**: 31–36.

[4] Pletnev VZ, Weeks CM, Duax WL. Rational proteomics II: electrostatic nature of cofactor preference in the short-chain oxidoreductase (SCOR) enzyme family. *Proteins*, 2004, **57**(2): 294–301.

[5] Yvonne K, Bengt P. Prediction of coenzyme specificity in dehydrogenases/reductases. A hidden Markov model-based method and its application on complete genomes. *FEBS J*, 2006, **273**(6): 1177–1184.

[6] Rodríguez-Zavala JS. Enhancement of coenzyme binding by a single point mutation at the coenzyme binding domain of *E. coli* lactaldehyde dehydrogenase. *Protein Science*, 2008, **17**: 563–570.

[7] Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res*, 2000, **28**: 304–305.

[8] Bairoch A, Apweiler R, Wu CH, *et al*. The universal protein resource (UniProt). *Nucleic Acids Res*, 2005, **3**: 154–159.

[9] Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 1997, **25**: 3389–3402.

[10] Caffrey DR, Somaroo S, Hughes JD, *et al*. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci*, 2004, **13**: 190–202.

[11] Jia J, Yang L, Zhang ZZ. EHPred: an SVM-based method for epoxide hydrolases recognition and classification. *J Zhejiang Univ Scicence B*, 2006, **7**(1): 1–6.

[12] Chou KC. Prediction of protein cellular attributes using pseudo-amino-acid-composition. *PROTEINS*: *Struct Func Genetics*, 2001, **43**: 246–255.

[13] Hellberg S, Sjostrom M, Skagerberg B, *et al*. Peptide quantitative structure-activity relationships, a multivariate approach. *J Med Chem*, 1987, **30**: 1126–1135.

[14] Andersson PM, Sjostrom M, Lundstedt T. Preprocessing peptide sequences for multivariate sequence-property analysis. *Chemometr Intell Lab Sys*, 1998, **42**: 41–50.

[15] Park KJ, Kanehisa M. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, 2003, **19**: 1656–1663.

[16] Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett*, 2006, **27**: 861–874.

[17] Inamdar NM, Ehrlich KC, Ehrlich M, *et al*. Data mining in bioinformatics using *Weka*. *Bioinformatics*, 2004, **20**: 2479–2481.

[18] Chou KC, Shen HB. Review: Recent progresses in protein subcellular location prediction. *Anal Biochem*. 2007, **370**: 1–16.

[19] Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, 2005, **21**: 10–19.

[20] Feng ZP. Zhang CT. Prediction of membrane protein types based on the hydrophobic index of amino acids. *J Protein Chem,* 2000, **19**: 269–275.

[21] Horne DS. Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. *Biopolymers*, 1988, **27**: 451–477.

[22] Sokal RR, Thomson BA. Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. *Am J Phys Anthropol*, 2006, **129**: 121–131.

[23] Cui J, Han LY, Lin HH, *et al*. Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties. *Mol Immunol*. 2007, **44**: 866–877.

[24] Chou KC. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem Biophy Res Comm*, 2000, **278**: 477–483.

[25] Hirano J, Miyamoto K, Hiromichi O. Purification and characterization of the alcohol dehydrogenase with a broad substrate specificity originated from 2-phenylethanol-assimilating *Brevibacterium* sp. KU 1309. *J Biosci Bioeng*, 2005, **100**(3): 318–322.

[26] Bengt P, Yvonne K, Udo O, *et al*. Coenzyme-based functional assignments of short-chain dehydrogenases/ reductases (SDRs). *Chemico-Biological Interactions,* 2003, **144**: 271–278.