DOI:10.16866/j.com.app.chem2006.02.018



# 基于两层分类器的半胱氨酸氧化还原状态预测方法

宋江宁[1,2], 李炜疆[1,2], 须文波[3]
(1. 江南大学工业生物技术教育部重点实验室, 江苏, 无锡, 214036; 2. 江南大学生物工程学院, 江苏, 无锡, 214036; 3. 江南大学信息工程学院, 江苏, 无锡, 214036)

**摘要**: 提出了两层混合分类器来预测蛋白质半胱氨酸氧化还原状态,第一层总体线性分类器利用氨基酸百分含量作为输入信息,第二层局部SVM分类器利用半胱氨酸周围局部序列作为输入信息。以2002年4月份的PISCES culled PDB数据库中的639条蛋白质多肽链作为研究对象,共含有584条二硫键,2 904个半胱氨酸。经严格的折叠刀方法检验,预测半胱氨酸的氧化还原状态准确率最高可达84.1%(半胱氨酸水平)和80.1%(蛋白质水平)。结果表明这种将蛋白质总体信息与局部上下文序列信息结合起来构建的两层混和分类器具有较高的预测准确率。研究结果也表明总体氨基酸百分含量和半胱氨酸周围局部序列都携带有二硫键形成的相关信息,暗示了半胱氨酸是否形成二硫键不但取决于蛋白质全局的结构信息同时也受到局部序列信息的影响。

**关键词**: 二硫键; 半胱氨酸; 氨基酸百分含量; 支持向量机; 协同性
**中图分类号**: Q 61
**文献标识码**: A
**文章编号**: 1001-4160(2006)02-177-182

# A novel prediction method of the oxidization state of cysteines in proteins based on two-stage classifier

SONG JiangNing[1,2], LI WeiJiang[1,2] and XU WenBo[3]
(1. The Key Laboratory of Industrial Biotechnology, Ministry of Education, Southern Yangtze University, Wuxi, 214036, Jiangsu, China; 2. School of Biotechnology, Southern Yangtze University, Wuxi, 214036, Jiangsu, China; 3. School of Information Engineering, Southern Yangtze University, Wuxi, 214036, Jiangsu, China)

**Abstract**: A novel approach has been proposed to predict the disulfide bonding state of cysteines in proteins by constructing a two-stage classifier combining a first global linear discriminator based on their amino acid composition and a second local support vector machine classifier. The present study was based on the PISCES Culled PDB with 639 protein polypeptide chains and 584 disulfide bonds in April 2002. The overall prediction accuracy of this hybrid classifier for the disulfide bonding state of cysteines in proteins has scored 84.1% and 80.1%, when measured on cysteine and protein basis using the rigorous jack-knife procedure, respectively. The result demonstrates the applicability of this novel method and provides comparable prediction performance compared with existing methods for the prediction of the oxidation states of cysteines in proteins. It also indicates that whether cysteines should form disulfide bonds depends not only on the global structural features of proteins but also on the local sequence environment of proteins.

**Key words**: disulfide bond, cysteine, amino acid composition, support vector machine, cooperativity

Song JN, Li WJ and Xu WB. A novel prediction method of the oxidization state of cysteines in proteins based on two-stage classifier. Computers and Applied Chemistry, 2006, 23(2):177-182.

## 1 引言

二硫键(Disulfide Bond)普遍存在于原核和真核生物的蛋白质中,是一类在蛋白质多肽链的两个半胱氨酸(Cysteine)之间形成的共价键。将形成二硫键的两个CYS的状态称为氧化态(oxidized state),没有形成二硫键的CYS的状态则称为还原态(reduced state)或者硫醇态(thiol form)。由于有氧情况下硫醇不稳定,胞外环境中的蛋白质半胱氨酸容易氧化形成二硫键,而胞内环境中的蛋白质由于其周围的还原性环境影响,半胱氨酸可以保持其活性基团不被氧化为二硫键,所以二硫键这种结构单元常常存在于非还原性环境中。二硫键的形成对于稳定蛋白质的空间结构和保持其活性功能具有极其重要的影响,它的错误配对是影响蛋白质多肽链正确折叠的重要原因[1,2]。二硫键形成是蛋白质折叠过程中的重要步骤,其形成动力学影响着蛋白质折叠的速率和途径[3]。二硫键除了对于单分子折叠反应具有动力学和热力

**收修改稿日期**: 2005-10-18
**基金资助**: 教育部工业生物技术重点实验室访问学者基金资助课题(200102)
**作者简介**: 宋江宁(1978—),男,山东,博士,研究方向:生物信息学.

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

学上的影响外,它还可以调节细胞内环境中分子间的接触从而可以影响蛋白质的分泌情况[4]。研究形成二硫键的蛋白质序列特征,找出与二硫键形成有关联的某些结构信息,对于蛋白质工程和人工药物分子设计都有着积极而重要的意义[5]。

在蛋白质折叠预测中,确定二硫键的形成位置可在很大程度上减少对于蛋白质构象空间的搜索[6],因而二硫键的准确预测会有许多潜在的重要应用,例如在蛋白质工程中引入人工二硫键以增加蛋白质结构的稳定性和帮助提高蛋白质三维空间结构的预测准确率。

在蛋白质二级结构中,两个氧化态 CYS 由于相互之间的空间接触(spatial contact)而形成二硫键,这基本上属于一种空间结构特征[5]。二硫键形成与什么信息有关,如何准确有效地预测其形成,这是一个有趣有意义的课题。基本上,二硫键形成预测方法可以分解为两步。首先,从一维氨基酸序列出发来预测蛋白质中的每个 CYS 的二硫键形成状态(disulfide bonding state of cysteine),或称为氧化还原状态(redox state of cysteine),这是一个典型的二元分类问题;其次,第二步是从那些预测为氧化态的候选 CYS 中确定二硫键的位置,即预测究竟是哪两个氧化态 CYS 之间形成了二硫键[6]。

本文的兴趣在于二硫键预测的第一步,即预测 CYS 的氧化还原状态。这方面的研究近年来逐渐受到关注,采用的研究方法主要有:

(1)统计分析(Statistical Analysis),Fiser 等人利用 CYS 周围局部序列信息来预测 CYS 的二硫键形成状态,达到了 71% 的准确率。Mucchielli-Giorgi 及其合作者使用具有相似氨基酸含量的蛋白质数据集训练的逻辑函数来预测半胱氨酸的二硫键形成状态,达到接近 84% 的预测成功率[7,8];

(2)多序列联配(Multiple Sequence Alignment),Fiser 和 Simon 利用多序列联配预测氧化态和硫醇态 CYS,准确率达到了 82% 以上[9];

(3)神经网络(Neural Network,NN),Muskal 等人利用 CYS 氨基酸两侧局部的氨基酸序列作为输入,总体预测准确率达到 80%[10],Fariselli 等人加上其他的进化信息进行预测,获得了更高的准确率[11];

(4)支持向量机(Support Vector Machine,SVM),Ceroni 等人在 Frasconi 等人[12]研究的基础上,使用相同的蛋白质二硫键数据集,对 SVM 方法进行了优化,总体准确率提高到 85% 左右[13];

(5)隐马尔科夫链(Hidden Markov Model,HMM),最近 Martelli 等人结合神经网络和隐马氏链方法构建了一种新的分类器,在蛋白质水平和 CYS 水平上的预测准确率分别是 88% 和 84%[14,15]。

本文提出了一种预测蛋白质半胱氨酸氧化还原状态的方法,成功构建了一个两层混合分类器—第一层分类器是一个建立在氨基酸百分含量上的总体二元线性分类器,利用蛋白质总体氨基酸百分含量作为输入信息,第二层分类器是一个利用中心 CYS 周围氨基酸序列作为输入信息的局部二元 SVM 分类器,结果表明这种混合分类器具有较高的预测准确率。

## 2 材料与方法

### 2.1 材料

非同源蛋白质结构数据出自 Wang 和 Dunbrack 等建立的 PISCES Culled PDB 数据库[16](2002 年 4 月),可过网址 http://www.fccc.edu/research/labs/dunbrack/pisces/culledpdb.html 上获得。由 PISCES 服务器获取 PDB 数据库中分辨率小于 0.25 nm,且序列同一性(sequence identity)低于 25% 的肽链结构数据,将这些结构数据集称为 PDB2525。在此基础上,选取序列长度大于 50 个氨基酸并且蛋白质结构分辨率 <2.5? 的 X 射线衍射数据构成出发数据库。

在 PDB 文件中,一般使用 SSBOND 记录来说明在肽链中出现的二硫键,我们在整理二硫键数据集的时候,用 FORTRAN 程序来读取二硫键形成(包括)的这部分信息,并读出相应的 PDB 文件名和氨基酸序列。由于二硫键有链内和链间二硫键之分,在此我们只统计链内二硫键(即在同一条蛋白质多肽链内形成的二硫键)的情况。

最终建成的数据库总共有 639 个 PDB 文件,总残基数是157 815个,总 CYS 个数为2 904。其中含有二硫键的有 218 个 PDB 文件(PDB 码列于表 1 中),总共有 584 条蛋白质链内二硫键。

根据蛋白质是否含有链内二硫键,可将数据分为两个集合,分别称为 OXICYS 集和 REDCYS 集。REDCYS 集中的蛋白质不含有二硫键,序列中的 CYS 都以还原态形式存在;OXICYS 集中的每个蛋白质至少含有一条二硫键。数据库中总共有 639 个蛋白质,其中 218 个 PDB 属于 OXICYS 集,421 个 PDB 属于 REDCYS 集。

本文所用数据库中蛋白质二硫键和 CYS 的分布情况(蛋白质序列分别依据其所含的二硫键和 CYS 数目进行分类),如图 1 和图 2 所示。从图 1 可以看出,含有 5 条以上二硫键数目的 OXICYS 集蛋白质序列只占到总数的大约 9%(20 / 218),大多数序列形成二硫键的数目不超过 5 条。与此相类似,大多数蛋白质序列中含有的 CYS 数目也在 10 个以内,数据库中 CYS 数目超过 10 的蛋白质序列只占 5% 的比例(33 / 639)(图 2)。

### 2.2 方法

#### 2.2.1 半胱氨酸氧化还原状态的协同性

218 个含有二硫键的 OXICYS 集蛋白质有1 316个半胱氨酸,其中有1 168个半胱氨酸参与形成了链内二硫键,亦即几乎所有的 OXICYS 蛋白质中的 CYS 以氧化态形式存在(在本文使用的数据集中,这一比例为 1 168/1 316 = 88.8%),而其余 421 个 REDCYS 蛋白质中的半胱氨酸为还原态。这样,蛋白质中 CYS 的氧化还原状态表现出一种明显的协同性现象(Cooperativity of the oxidization of cysteines in proteins),这是个很有趣的现象。这一现象不是巧合,这种

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

明显的协同性现象仅仅通过 CYS 周围的局部序列还不足以解释清楚。这一协同性是与蛋白质结构有关的一种全局特征,可能存在一些全局信息来说明这一现象。

表 1　数据库中含有二硫键的 218 个 PDB 码

Table 1　The PDB codes of 218 proteins containing disulfide bonds in the dataset.

| 153L | 1A6WL | 1A6WH | 1AAZA | 1ABRB | 1AC5 | 1AGQA | 1AHO | 1AIR | 1AISA |
|---|---|---|---|---|---|---|---|---|---|
| 1ALKA | 1ALU | 1AMP | 1AOCA | 1AOZA | 1APA | 1APYA | 1APYB | 1AQB | 1AQZA |
| 1ARB | 1ARU | 1ATLA | 1AU1B | 1AUK | 1AV4 | 1BEBA | 1BEO | 1BGC | 1BGP |
| 1BHP | 1BNDA | 1BNDB | 1BOVA | 1BPI | 1BTL | 1CELA | 1CEX | 1CFB | 1CGT |
| 1CNSA | 1CNV | 1CPO | 1CTN | 1CVL | 1DANL | 1DANH | 1DANT | 1DANU | 1DDT |
| 1DPE | 1EAGA | 1ECEA | 1ECY | 1EDMB | 1EPTA | 1EPTC | 1ESC | 1EXTA | 1EZM |
| 1FLEE | 1FLEI | 1FUS | 1FVKA | 1FXD | 1G3P | 1GAI | 1GAL | 1GEN | 1GOF |
| 1HCGA | 1HCGB | 1HIAA | 1HIAB | 1HIAI | 1HOE | 1HPLA | 1HSBA | 1HSBB | 1HSSA |
| 1HTRB | 1HUCA | 1HUCB | 1HXN | 1HYP | 1IAE | 1IDK | 1ILR1 | 1IMBA | 1IVYA |
| 1JER | 1JETA | 1JFRA | 1JMCA | 1JPC | 1KLO | 1KPTA | 1KSIA | 1KTE | 1KUH |
| 1KVEA | 1KVEB | 1LBEA | 1LBU | 1LIT | 1LKI | 1LPBA | 1LPBB | 1LST | 1LT5D |
| 1LTSD | 1MHLA | 1MHLC | 1MPP | 1MUP | 1MZM | 1NEU | 1NNC | 1NOYA | 1NSCA |
| 1NWPA | 1OBR | 1ONC | 1OVA | 1PGS | 1POA | 1POC | 1PPN | 1PYTB | 1PYTC |
| 1PYTD | 1RCB | 1RFS | 1RGEA | 1RIE | 1RMG | 1SFP | 1SMD | 1SMNA | 1SMPI |
| 1SRA | 1SVB | 1TABE | 1TABI | 1TCA | 1TDE | 1TF4A | 1TFE | 1TGSZ | 1TGSI |
| 1TGXA | 1THG | 1THV | 1TIID | 1TML | 1TN3 | 1TVDA | 1UKZ | 1UMAH | 1VCAA |
| 1VMOA | 1WBA | 1WHTA | 1XJO | 1XSOA | 1YAIA | 1ZXQ | 2AAA | 2ACK | 2AMG |
| 2AYH | 2BBKH | 2BBKL | 2CBP | 2CTC | 2DNJA | 2ENG | 2ERL | 2GMFA | 2HLCA |
| 2ILK | 2LIV | 2MCM | 2MPRA | 2MSBA | 2MTAH | 2MTAL | 2OVO | 2PKAA | 2PKAB |
| 2PSPA | 2RHE | 2SAS | 2SGA | 2SICI | 2SIL | 2TGI | 2TRXA | 2VPFA | 2WEA |
| 3CD4 | 3EBX | 3FRUA | 3FRUB | 3GRS | 3LADA | 3LZT | 3PTE | 3SEB | 3TGL |
| 4AAHA | 4AAHB | 4HTCH | 4HTCI | 5PTP | 7RSA | 8FABA | 8FABB | | |

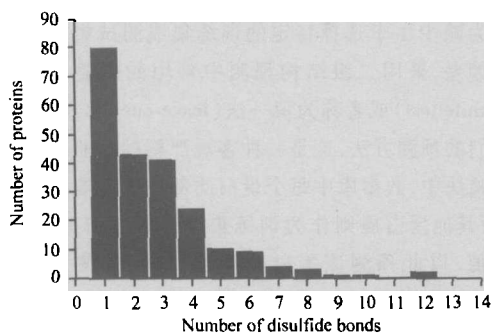


Fig. 1　Distribution of disulfide bridges per protein sequences in the dataset. Protein chains are grouped according to the number of disulfide bonds.

图 1　数据库中 OXICYS 集蛋白质二硫键的分布情况,依据蛋白质序列中的二硫键数目进行分类

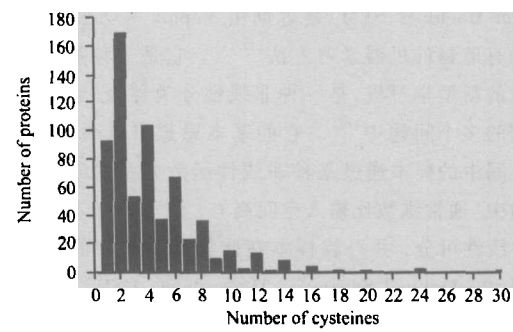


Fig. 2　Distribution of cysteines per protein sequences in the dataset. Protein chains are grouped according to the number of cysteines.

图 2　数据库中所有蛋白质 CYS 的分布情况,依据蛋白质序列中的 CYS 数目进行分类

### 2.2.2　第一层分类器—总体二元线性分类器

第一层分类器是一个全局的二元线性分类器,利用蛋白质总体氨基酸百分含量作为输入信息。对数据集中的某一个蛋白质 $k$ 而言,定义特征参数 $Q_k$ 如下:

$$Q_k = \begin{cases} +1, \text{如果蛋白质 } k \text{ 属于 OXICYS 集,} \\ -1, \text{如果蛋白质 } k \text{ 属于 REDCYS 集.} \end{cases} \quad (1)$$

由蛋白质序列的氨基酸组分 $p_a^{(k)}$ 出发得到蛋白质 $k$ 的特征参数 $Q_k$,采用简单的线性函数 $p_a^{(k)}$ 来约化 $Q_k$,即

$$Q_k = \sum_a v_a p_a^{(k)} \quad (2)$$

$a$ 代表氨基酸,式(2)对所有的 20 种氨基酸求和得到 $Q_k$。参数 $v_a$ 对所有的蛋白质而言为一常数。为求出数据集中蛋白质的最佳 $v_a$ 值,对下式最小化:

$$Z = \sum_k \left( Q_k - \sum_a v_a p_a^{(k)} \right)^2 \quad (3)$$

对氨基酸 $b$ 令 $\partial Z/\partial v_b = 0$,得到

$$\sum_a \left( \sum_k p_a^{(k)} p_b^{(k)} \right) v_a = \sum_k Q_k p_b^{(k)} \quad (4)$$

在此,对训练数据集中的所有的蛋白质序列求和。式(4),得到 20 种氨基酸的最佳参数值 $v_a$。据此计算氨基酸百分含量 $p_a$ 的某一蛋白质序列的 $Q$ 值:

$$Q = \sum_a v_a p_a \quad (5)$$

预测时,某一蛋白质如果属于 OXICYS 集蛋白质,那么 $Q$ 值假定为 +1,如果属于 REDCYS 集蛋白质,$Q$ 值假定为 -1。这里 $Q$ 值作为蛋白质分类的特征参数。为检验使用 $Q$ 值的合理性,计算累积分布:

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

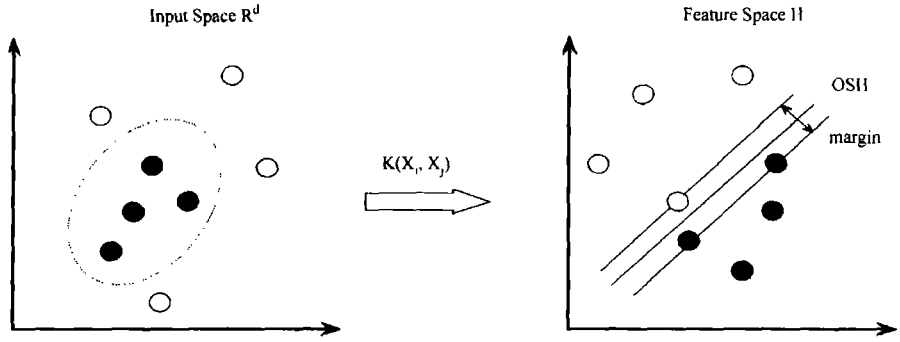


Fig. 3 Basic idea of SVM application for pattern recognition. Two classes denoted by circles and disks, respectively.

图3 支持向量机方法用于模式识别的基本思想,两类样本分别用空心和实心圆圈表示

$$F_{\text{OXICYS}}(Q_c) = \frac{\text{The number of OXICYS proteins with } Q \geq Q_c}{\text{The number of all OXICYS proteins}} \quad (6)$$

$$F_{\text{REDCYS}}(Q_c) = \frac{\text{The number of REDCYS proteins with } Q \geq Q_c}{\text{The number of all REDCYS proteins}} \quad (7)$$

在此,$Q_c$ 为两类蛋白质分类的临界值。

### 2.2.3 第二层分类器—局部二元 SVM 分类器

第二层分类器是一个局部二元 SVM 分类器,利用目标 CYS 周围的氨基酸序列作为输入信息。支持向量机(Support Vector Machines,SVM)是近期由 Vapnik 等人提出的一种有效的有监督的机器学习方法[17,18]。它是一种基于统计学习理论的新型学习机,是一种非线性分类器,已经被用于模式识别的多个问题中[19]。它的基本思想可以表述如下:将输入空间中的样本通过某种非线性函数关系映射到一个特征空间中(通常维数比输入空间高),使两类样本在此特征空间中线性可分,并寻找样本在此特征空间中的最优分割超平面(the Optimal Separating Hyperplane, OSH),使两类样本具有最大的分开距离,从而将两类样本成功地分离(图2)。

对二元分类问题(本文为形成二硫键的 CYS 和不形成二硫键的 CYS 两种状态),假设已有一系列样本,即一系列输入向量 $x_i \in R^d (i = 1,2,\cdots,N)$,其相应标识 $y_i \in \{+1, -1\}$ $(i = 1,\cdots,N)$,+1 对应正类样本(氧化态 CYS),-1 对应负类样本(还原态 CYS)。本文中,输入向量维数为 20,每个输入单元为一个围绕中心 CYS 残基的序列片断滑动窗,窗口长度为 $l = 2k + 1$($k$ 表示每个中心 CYS 周围序列从 N 端到 C 端的氨基酸残基数目,$k = \cdots, 7, 8, 9, 10, \cdots$)。将 20 种氨基酸组成的输入序列转换为数字形式编码,由 0 和 1 所组成(Ala = 100000…000, Cys = 010000…000, …, Tyr = 000000…001)。

映射是通过采用核函数 $K(x_i, x_j)$ 技术实现的,核函数定义了特征空间的内积。主要有以下两种常用的核函数:

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^d \quad (8)$$

$$K(x_i, x_j) = \exp(-r \| x_i - x_j \|^2) \quad (9)$$

在此,式(8)为 $d$ 次方的多项式核函数,在 $d = 1$ 时即为线性函数。式(9)为径向基核函数(RBF Kernel Function),有一个可调参数 $r$。

对给定的数据集而言,只需要选择核函数类型和正则化参数 $C$ 来指定 SVM。在本文中,我们最终选择多项式核函数来训练 SVM。多项式核函数定义为:$K(x_i, x_j) = (x_i \cdot x_j + 1)^d$,正则化参数 $C$ 采用默认数值,$d = 7,8,9$。

支持向量机软件包来自于 SVM$^{\text{light}}$,可以通过以下网址下载得到:

http://download.joachims.org/svm_light/current/svm_light_windows.zip,这是一个 SVM 二元分类器的 C 语言程序[20,21]。

### 2.2.4 预测准确率检验方法

为减少由于选择特定的训练集或测试集而导致的预测结果偏差,采用二级结构预测中常用的折叠刀检验法(the jack-knife test)或者称为留一法(leave-one-out procedure)来评价我们的预测方法,这是一种客观严格的检验方法。在折叠刀检验法中,数据库中每个蛋白质都被依次取出作为测试蛋白,而其他蛋白质则作为训练集以计算预测时所需要的 va 参数值,以此预测该蛋白质属于 OXICYS 集还是 REDCYS 集。预测效果分别用 CYS 水平和蛋白质水平上的预测准确率表示。

我们采用总预测准确率($Q_2$)、蛋白质水平上的预测准确率($Q2_{\text{prot}}$)、Matthew 相关系数(MCC)作为预测方法的评价指标。

定义 $n_{xy}$ 为预测成 $x$ 集合蛋白但实际上却为 $y$ 集合蛋白的的蛋白质个数,在此 $x, y = o$(OXICYS),或者 $r$(REDCYS)。那么预测结果可以通过下列方程来评价:

$$Q2 = P/N = \frac{n_{oo} + n_{rr}}{n_{oo} + n_{or} + n_{rr} + n_{ro}} \quad (10)$$

其中 $P$ 为状态预测正确的 CYS 的数目,$N$ 是 CYS 的总数目。

预测效果评价的另一种方法是 Matthew 相关系数法[22](Matthew's Correlation Coefficient, MCC),如下式所示:

$$\text{MCC}(s) = \frac{p(s)n(s) - u(s)o(s)}{\sqrt{(p(s)+u(s))(p(s)+o(s))(n(s)+u(s))(n(s)+o(s))}} \quad (11)$$

$p(s)$ 和 $n(s)$ 分别指正确预测是(correct prediction)的数目和正确预测不是(correctly rejected assignments)的数目,而 $u(s)$ 和 $o(s)$ 指估计过剩(over-prediction)和估计不足(under-pre-

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

diction)的数目。通常 MCC 值越大,表明预测结果可靠性也越好。

两类蛋白质集合 $s$(OXICYS 集或 REDCYS 集)的预测结果可通过下式来评价:

$$Q_{oxi} = \frac{n_{oo}}{n_{oo} + n_{or}}, Q_{red} = \frac{n_{rr}}{n_{rr} + n_{ro}} \quad (12)$$

其中 $Q_{oxi}$ 和 $Q_{red}$ 分别是 OXICYS 集和 REDCYS 集的预测准确率。

蛋白质水平上的整体预测准确率为:

$$Q2_{prot} = \frac{P_p}{N_p} \quad (13)$$

$P_p$ 表示序列内所有 CYS 的状态都被正确预测出的蛋白质数目,$N_p$ 是总蛋白质数目。

## 3　结果与讨论

### 3.1　20 种氨基酸对二硫键形成的贡献

对于 20 种氨基酸,可通过式(4)算出这 20 种氨基酸经过数据集训练后的参数值 $v_a$,由此得到它们对于二硫键形成的贡献权重(图 4)。由图 4 可以看出,半胱氨酸(Cysteine,C)、天冬酰胺(Asparagine,N)、丝氨酸(Serine,S)、苏氨酸(Threonine,T)对于二硫键形成有强烈的正相关性,即这些蛋白质序列有这些氨基酸的存在会有强烈的形成二硫键的倾向;而谷氨酸(Glutamate,E)、组氨酸(Histidine,H)、亮氨酸(Leucine,L)、蛋氨酸(Methionine,M)、缬氨酸(Valine,V)和精氨酸(Arginine,R)则表现出强烈的二硫键形成负相关性。这些结果与以前的研究基本一致[8,9]。

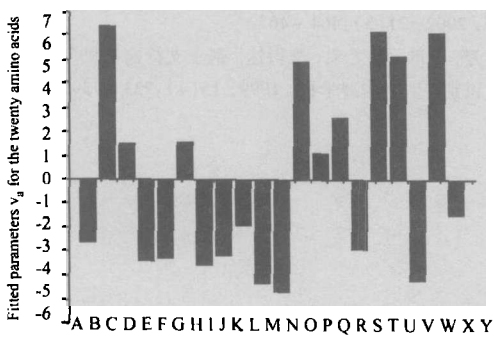


Fig. 4　Amino acid contribution to disulfide bond formation. Calculated $v_a$ values can represent the propensity to form disulfide bond for the twenty amino acid residues. Bars above the midline indicate a propensity to disulfide bond formation, and those below the midline are inclined to non-disulfide bond formation.

图 4　20 种氨基酸对于二硫键形成的贡献;参数 $v_a$ 值可表示 20 种氨基酸形成二硫键的倾向;中线上面的条形框表示此氨基酸对于二硫键形成有正相关性,中线下面的条形框表示此氨基酸对二硫键形成有负相关性

### 3.2　第二层 SVM 分类器的参数优化

对于第二层 SVM 分类器而言,我们需要选择合适的核函数类型及参数,以及正则化参数 C 和局部对称输入序列滑动窗长度 $l = 2k + 1$ ($k = \cdots, 7, 8, 9, 10, \cdots$)。最适核函数类型和正则化参数 C 的选择对提高总体预测准确率有很重要的影响。最优参数根据系统预测效果而定。

我们预先进行了一系列测试,改变滑动窗口的长度 $l$($l$ 从 13 变化到 21,相应地,$k$ 从 6 变化到 10)根据预测准确率高低来确定最优的核函数类型,最合适的滑动窗口长度 $l$。我们比较了线性,多项式和径向基核函数对预测结果的影响。最终预测中使用的都是那些能产生最佳预测结果的最优参数和核函数。最后,我们选择 $d = 8$ 的多项式核函数来进行 SVM 的训练和预测,局部序列滑动窗口长度 $l = 21$。

### 3.3　两层混合分类器的预测结果

折叠刀检验法得到的预测结果如表 2 所示。当选择自然阈值 $Q_c = 0$ 时,预测结果为:总体预测准确率为 $Q2 = 83.3\%$(CYS 水平),$Q2_{prot} = 79.4\%$(蛋白质水平),其中氧化态 CYS 和还原态 CYS 预测准确率分别为 $Q_{oxi} = 85.7\%$,$Q_{red} = 78.9\%$,相关系数为 MCC = 59.6%。采用动态调整阈值的方法,可在一定程度上进一步提高预测准确率。当选择阈值 $Q_c = -0.1$ 时,则得到最高的预测准确率:总体预测准确率为 $Q2 = 84.1\%$(CYS 水平),$Q2_{prot} = 80.1\%$(蛋白质水平),其中氧化态 CYS 和还原态 CYS 预测准确率分别为 $Q_{oxi} = 87.8\%$, and $Q_{red} = 77.8\%$,相关系数为 MCC = 62.2%。

表 2　两层混合分类器预测系统的预测准确率

Table 2　Prediction accuracy (%) of the two-stage hybrid classifier prediction system by the jack-knife test.

| Method | Prediction accuracy (%) | | | | | |
|---|---|---|---|---|---|---|
| | $Q_c$ | MCC | $Q_{oxi}$ | $Q_{red}$ | $Q2$ | $Q2_{prot}$ |
| Global linear classifier + Local SVM classifier | -0.2 | 61.6 | 89.5 | 74.2 | 83.2 | 79.5 |
| | -0.1 | 62.2 | 87.8 | 77.8 | 84.1 | 80.1 |
| | 0 | 59.6 | 85.7 | 78.9 | 83.3 | 79.4 |
| | 0.1 | 58.3 | 84.2 | 79.9 | 82.9 | 78.5 |
| | 0.2 | 55.7 | 81.7 | 80.7 | 81.9 | 77.9 |

以上预测结果表明这种将蛋白质总体信息与局部上下文序列信息结合起来构建的两层混和分类器具有较高的预测准确率。研究结果也表明总体氨基酸百分含量和半胱氨酸周围局部序列都携带有二硫键形成的相关信息,暗示了半胱氨酸是否形成二硫键不但取决于蛋白质全局的结构信息同样也受到其局部序列信息的影响。

## 4　结论

本文提出了一种新的两层混合分类器方法来预测蛋白质半胱氨酸氧化还原状态:第一层分类器利用总体氨基酸百分含量作为输入信息,第二层分类器利用半胱氨酸局部序列作为输入信息。结果表明,我们建立的这种新预测系统具有较高的预测准确率。经严格的折叠刀方法检验,预测半胱氨酸的氧化还原状态准确率最高可达 84.1%(半胱氨酸水平)和 80.1%(蛋白质水平)。研究结果表明半胱氨酸是否形成二硫键不但取决于蛋白质全局的结构信息同样也受到局部序列信息的影响,也说明使用这种两层混合分类算法能够有

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

效地提高蛋白质半胱氨酸氧化还原状态的预测准确率。

References


1 Wedemeyer WJ, Welker E, Narayan M and Scheraga HA. Disulfide bonds and protein folding. Biochemistry, 2000, 39(15):4207 - 4216.

2 Harrison PM and Sternberg MJE. Analysis and Classification of disulphide connectivity in proteins. J Mol Biol, 1994, 244:448 - 463.

3 Wittrup KD. Disulfide bond formation and eukaryotic secretory productivity. Current Opinion in Biotechnolgoy, 1995, 6:203 - 208.

4 Abkevich VI and Shakhnovich EI. What can disulfide bonds tell us about protein energetics, function and folding: simulations and bioinformatics analysis. J Mol Biol, 2000, 300:975 - 985.

5 Song JN and Li WJ. Analysis of sequence features of disulfide-bonds in roteins. Journal of Wuxi University of Light Industry, 2002, 21(5):464 - 467.

6 Fariselli P and Casadio R. Prediction of disulfide connectivity in proteins. Bioinformatics, 2001, 17(10):957 - 964.

7 Mucchielli-Giorgi MH, Hazout S and Tuffery P. Predicting the disulfide bonding state of cysteines using protein descriptors. Proteins, 2002, 46:243 - 249.

8 Fiser A, Cserzo M, Tudos E and Simon I. Different sequence environment of cysteines and half cystines in proteins: Application to predict disulfide forming residues. FEBS Lett, 1992, 302:117 - 120.

9 Fiser A and Simon I. Predicting the oxidation state of cysteines by multiple sequence alignment. Bioinformatics, 2000, 16:251 - 256.

10 Muskal S, Holbrook S and KIM S. Prediction of the disulfide-bonding state of cysteine in proteins. Protein Eng, 1990, 3:667 - 672.

11 Fariselli P, Riccobelli P and Casadio R. Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. Proteins, 1999, 36:340 - 346.

12 Frasconi P, Passerini A and Vullo A. A two stage SVM architecture for predicting the disulfide bonding state of cysteines. In Proceedings of IEEE Neural Network for signal processing conference. IEEE Press, 2002.

13 Ceroni A, Frasconi P, Passerini A and Vullo A. Predicting the disulfide bonding state of cysteines with combinations of kernel machines. Journal of VLSI Signal Processing, 2003, 35:287 - 295.

14 Martelli PL, Fariselli P, Malaguti L and Casadio R. Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks. Protein Eng, 2002, 15:951 - 953.

15 Martelli PL, Fariselli P, Malaguti L and Casadio R. Prediction of the disulfide bonding state of cysteines in proteins at 88% accuracy. Protein Sci, 2002, 11:2735 - 2739.

16 Wang G and Dunbrack RLJr. PISCES: a protein sequence culling server. Bioinformatics, 2002, 19:1589 - 1591.

17 Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

18 Vapnik V. Statistical Learning Theory. New York: John Wiley and Sons, Inc., 1998.

19 Wen F, Lu X, Sun ZR and Li YD. Splice sites prediction using support vector machine. Acta Biophysica Sinica, 1999, 15(4):733 - 739.

20 Joachims T. Making large-scale SVM learning practical. In: Advances in Kernel Methods-Support Vector Learning. Cambridge, MA: MIT Press, 1999.

21 Joachims T. Learning to classify text using support vector machine. Dissertation, Kluwer, 2002.

22 Matthews BW. Comparison of predicted and observed secondary structure of T4 phage lysozyme. Biophys Acta, 1975, 405:442 - 451.


附中文参考文献


5 宋江宁，李炜疆. 蛋白质二硫键的分布特征. 无锡轻工大学学报，2002，21(5):464 - 467.

19 闻芳，卢欣，孙之荣，李衍达. 基于支持向量机(SVM)的剪接位点识别. 生物物理学报，1999，15(4):733 - 739.


?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net