Two Multi-classification Strategies Used on SVM to Predict Protein Structural Classes by Using Auto Covariance

Jiang Wu1,2, Yi-Zhou Li1, Meng-Long Li1*, Le-Zheng Yu1

1 (College of Chemistry, Sichuan University, Chengdu 610064, China)
2 (Department of Information Technology, Yulin University, Yulin 719000, China)

Received 28 July 2009 / Revised 1 September 2009 / Accepted 4 September 2009

Abstract: Machine learning methods play the very important role in protein secondary structure prediction and other related works. On condition of a certain approach, the prediction qualities mostly depend on the ways of representing protein sequences into numeric features. In this paper, two Support Vector Machine (SVM) multi-classification strategies, “one-against-one” (1-a-1) and “one-against-all” (1-a-a), were used in protein structural classes identification. Auto covariance (AC), which transforms the physicochemical properties of the amino acids of the proteins into a data matrix, focuses on the neighboring effects and the interactions between residues in protein sequences. “1-a-1” approach was used on SVM to predict protein structural classes and obtained very promising overall accuracy 90.69% by Jackknife test. It was more than 10% higher than the accuracy obtained by using “1-a-a”. Experimental results led to the finding that the SVM predictor constructed by “1-a-1” can avoid the appearance of biased prediction accuracy. This current method, using the protein primary sequence information described by auto covariance (AC) and “1-a-1” approach on SVM, should play an important complementary role in other related applications.

Key words: support vector machine, protein structural class, multi-classification, auto covariance.

1 Introduction

The structural class is one of the important features of a protein, and it is also a key step toward researching their three-dimensional structure and function. With the rapid increase of the gap between the number of known protein sequences and the number of known structures, it is badly needed to build a computational system for determining the protein structural class (Chen et al., 2006). For this step, machine learning and pattern recognition approaches have been particularly successful (Baldi et al., 1999). Using machine learning and artificial intelligence techniques to predict protein structural class and other related classification problems of protein mostly depended on protein sequence. Therefore, how to represent the amino acid residues into numerical values is the most important step. Some methods have been proposed to predict the structural classes of protein based on their amino acid (AA) compositions (Bahar et al., 1997; Zhou and Assa-Munt, 2001), but the representations based on AA composition alone will certainly lose many important patterns associated with the sequence order. Predictor based on pseudo-amino acid composition (PseAA) defined in Chou (2001) has been wildly used (Chou and Cai, 2003; Chou, 2005; Shen et al., 2006; Zhou et al., 2007). As addition of AA, PseAA takes the sequence order information of a protein into account. Other methods include predictor based on profiles (Guo et al., 2004), predictor based on function domain composition (Chou and Cai, 2004), predictor based on evolutionary information (Kaur and Raghava, 2004), predictors based on polypeptide composition and dipeptide components (Lin and Li, 2007; Luo et al., 2002), predictor based on the distance score of amino acid sequence (Zhong et al., 2007), and predictor based on the angles between amino acid triplets (Karci and Demir, 2008). Meanwhile, the machine learning approaches have obtained great achievements in protein secondary structure prediction (Baldi and Brunak, 1998). SVM (Vapnik, 1995, 1998) was designed for binary classification problems. And in most researches, “1-a-1” (Friedman, 1996) and “1-a-a” (Hsu and Lin, 2002 and references therein) strategies used on SVM were normally intended to solve multi-classification problems. In this article the two approaches and AC descriptors were used on SVM to predict protein structural classes. Auto cross covariance (ACC) developed by Wold et al.
(1993) has been widely used in classification problems (Doytchinova and Flower, 2007; Guo et al., 2006, 2008) and applied to describe the protein sequences in this article, it generates auto covariance (AC) and cross covariance (CC) variables. AC variables describe and explain the interactions about residues with a certain distance in between, it sufficiently has considered the adjacent effects and it possibly could dig out the patterns that run through entire sequences (Guo et al., 2008). The dataset used in this paper contains 204 proteins classified into four structural classes: α, β, α + β, α/β. First stage, “1-a-a” strategy intended for SVM was used to solve the multi-classification problem and α + β structural class subset got very bad prediction accuracy, lower than 60%. The reasons leading to the biased prediction accuracy were discussed in section 3. Second stage, “1-a-1” solution was used on SVM to predict protein structural classes and obtained 90.69% overall accuracy and no biased prediction accuracy appeared in structural class subsets. The prediction results obtained by this method and evaluated by Jackknife test exhibited improvement compared with several published results.

2 Materials and methods

2.1 Dataset

The dataset used in this article was constructed by Chou (1999). It contains 204 proteins which can be classified into four structural classes: 52 belong to α, 61 belong to β, 45 belong to α/β, and 46 belong to α + β according to the definition by Levitt and Chothia (1976). Each protein belongs to a subset $Sub_i$ ($i = 1, 2, \cdots, 4$) as shown in Table 1. Using symbol $\bigcup$ to express union operation, the structural dataset $S$ was formulated as Equation (1).

$$S = Sub_1 \bigcup Sub_2 \bigcup Sub_3 \bigcup Sub_4$$

(1)

2.2 Representing proteins by using AC

In this article, the following seven amino acid physicochemical properties were used in reflecting protein samples. They are hydrophobicity (Tanford, 1962), hydrophilicity (Hopp and Woods, 1981), volumes of side chains of amino acids (Krigbaum and Komoriya, 1979), polarity (Grantham, 1974), polarizability (Charton, M. and Charton, B.I., 1982), solvent-accessible surface area (SASA) (Rose et al., 1985), and net charge index (NCI) of side chains of amino acids (Zhou et al., 2006). The seven physicochemical properties of amino acids were used to predict protein-protein interactions (Guo et al., 2008). SVM needs same length instances for training. Auto covariance (AC) transformed the numerical instances into uniform inputs. To represent a protein sample with length $L$, the AC variables are calculated according to Equation (2), where $lag$ is the distance between residues, $j$ the physicochemical property of nature amino acids of the seven mentioned above, $i$ the position in the sequence of a protein sample $X$. In this paper a protein sample is represented as Equation (3), where $f_1, f_2, \cdots, f_{20}$ is the frequencies of the 20 native amino acids in the protein sequence.

$$AC_{lag,j} = \frac{1}{L-lag} \sum_{i=1}^{L-lag} \left( X_{i,j} - \frac{1}{L} \sum_{i=1}^{L} X_{i,j} \right)$$

(2)

$$P_{AC}^{lag} = [f_1, f_2 \cdots f_{20}, AC_{lag,j}]$$

(3)

In this way, a protein sequence was represented as a vector with dimensions $20 + lag \times j$, where, $j$ is the number of physicochemical property descriptors used in this article. In this section, a protein sequence was represented by 49 different vectors corresponding to different parameter $lag = (1, 2, \cdots, 49)$.

2.3 Multi-classification strategy on SVM

Decomposing a multi-classification problem to several binary problems is the main idea of “1-a-1” and “1-a-a” strategies on SVM. In “1-a-a” decomposition, $k$ classes multi-classification problem is transferred into $k$ binary classification problems. The $m - th (m = 1, 2, \cdots, k)$ classifier was built on the training data in which $m - th$ the class samples were trained as positive samples and all the others were trained as negative samples. “1-a-1” strategy, in which $k(k-1)/2$ classifiers are constructed, each of them was trained on two different classes, an input sample gets its prediction label through a voting process by all the classifiers, and it is classified belong to the class with maximum number of voters. Hsu and Lin (2002) compared these two strategies on SVM for multi-classification. The following parameters were used to evaluate the performance of the current method, where $T(i)$ the number of correctly predicted samples in class $i$, $Num(i)$ is the number of

### Table 1 The structural classes dataset consisting of 204 proteins are classified into four subsets

<table>
<thead>
<tr>
<th>Structural class</th>
<th>Subsets</th>
<th>Number of protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha/\beta$</td>
<td>$Sub_1$</td>
<td>45</td>
</tr>
<tr>
<td>$\alpha + \beta$</td>
<td>$Sub_2$</td>
<td>46</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$Sub_3$</td>
<td>52</td>
</tr>
<tr>
<td>$\beta$</td>
<td>$Sub_4$</td>
<td>61</td>
</tr>
<tr>
<td>Dataset</td>
<td></td>
<td>204</td>
</tr>
</tbody>
</table>

*See Equation (1).