A structured outputs method for predicting protein binding sites

Michael Hamilton HAMILTOM@CS.COLOSTATE.EDU
Asa Ben-Hur ASA@CS.COLOSTATE.EDU
Colorado State University, Computer Science Department, 1873 Campus Delivery, Fort Collins, CO 80523-1873

Keywords: structured outputs, kernel methods, calmodulin-binding, perceptron

Abstract

Protein-protein interactions have essential roles in nearly all biochemical processes. While high-throughput methods exist for experimentally identifying interaction partners, the task of determining binding site locations remains arduous. We consider the prediction of protein binding site location as an instance of the label sequence problem and outline a representation in the framework of structured outputs. Moreover, we compare kernel structured output methods with sliding window classifiers in the identification of calmodulin-binding sites.

1. Introduction

Determining the locations of binding sites within proteins can be viewed as a label sequence prediction problem. Predicting labels for sequences involves learning a mapping between input-space sequences $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and output-space sequences $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ such that $x_i \in \Sigma_x$, and $y_i \in \Sigma_y$. For binding site prediction, Σ_x is the 20 amino acid alphabet and $\Sigma_y = \{0, 1\}$, indicating whether the amino acid is part of the binding site (1) or not (0).

Calmodulin is a ubiquitous, highly-conserved calcium-binding protein found in eukaryotes (Vetter & Leclerc, 2003). Calmodulin binding sites share little sequence similarity; however, they have 3 common attributes: (1) each binding site consists of approximately 20 residues, (2) the binding sites are contiguous in sequence, (3) binding sites are slightly hydrophobic and have tendencies to form α-helices.

The size and contiguous properties, (1) and (2), are important as they reduce the prediction of calmodulin-binding site locations to a special case of the label sequence prediction problem. Whereas a sequence of length n can, in general, be associated with 2^n possible labellings, the contiguous sequence patterns of length k reduce the search spaces to $n - k + 1$ start positions of the binding site, assuming a single binding site.

2. Sliding Window Approach

A strategy for solving the label sequence learning problem uses sliding window classifiers (Dietterich, 2002). Each position in the sequence can be represented by a window of length k centered at that position. This reduces label sequence learning to a binary classification problem, each component in the label sequence is predicted by a binary classifier trained on fixed-length windows. The label sequence created in this process does not necessarily satisfy the constraints of the given problem: in our case for example, we want to predict a single binding site in the sequence. A step of post-processing is therefore necessary to determine the predicted label sequence.

Since our objective is to predict a single binding site that is contiguous in the protein sequence, the label space, $\mathcal{Y}$, for sequences of length n is composed of all length n 0-1 sequences that have k contiguous nonzero elements. In our application neighboring windows in the sequence overlap, and therefore their corresponding discriminant functions are likely to be correlated, and isolated positively-predicted windows are likely to be false-positives. Our strategy is therefore to average the discriminant function of the binary classifier over adjacent windows to form the window discriminant

$$F_r(\mathbf{x}, \mathbf{y}) = \frac{1}{2r+1} \sum_{i=c(\mathbf{y})-r}^{c(\mathbf{y})+r} f(w_i), \qquad (1)$$

where $\mathbf{y} \in \mathcal{Y}$ and $c(\mathbf{y})$ is the position of the center of the binding site implied by $\mathbf{y}$, $f(w_i)$ is the discrimi-

Preliminary work. Under review by the International Workshop on Mining and Learning with Graphs (MLG). Do not distribute.

nant value for a window centered at position i, and r is some positive integer indicating the half-width of the neighborhood to be considered (a value $r = 3$ was found to be optimal in our experiments). Label sequence inference is performed by

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} F_r(\mathbf{x}, \mathbf{y}). \tag{2}$$

3. Structured Outputs

Whereas sliding window methods first train a classifier over substrings of input sequences then generate compound labels, structured outputs combine these steps into a single learning framework (Altun et al., 2003). This is achieved by measuring the compatibility between input and possible output pairs.

Formally, let $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ be the set of annotated input-output pairs. For a given input-output pair, $(\mathbf{x}_i, \mathbf{y}_i)$, we define

$$F_{\mathbf{w}}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{w}, \phi(\mathbf{x}, \mathbf{y}) \rangle \,, \tag{3}$$

where $\phi(\mathbf{x}, \mathbf{y})$ embeds the pair into some input-output feature space. Equation 3 can be viewed as assessing the compatibility of a label $\mathbf{y}$ for an input sequence $\mathbf{x}$.

Inference in this framework is performed by finding the output $\hat{\mathbf{y}}$ that is most compatible with the input $\mathbf{x}$:

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} F_{\mathbf{w}}(\mathbf{x}, \mathbf{y}). \tag{4}$$

Since the number of elements in $\mathcal{Y}$ is linear in the length of the sequence, the computation of the argmax operation can be performed efficiently.

In order to employ kernels we use the dual form

$$F_\alpha(\mathbf{x}, \mathbf{y}) = \sum_i \sum_{\mathbf{y}'} \alpha_{i,\mathbf{y}'} \langle \Phi(\mathbf{x}_i, \mathbf{y}'), \Phi(\mathbf{x}, \mathbf{y}) \rangle \,, \tag{5}$$

which can be expressed using kernels as

$$F_\alpha(\mathbf{x}, \mathbf{y}) = \sum_i \sum_{\mathbf{y}'} \alpha_{i,\mathbf{y}'} k((\mathbf{x}_i, \mathbf{y}'), (\mathbf{x}, \mathbf{y})). \tag{6}$$

For the purpose of predicting binding sites we define

$$k((\mathbf{x}, \mathbf{y}), (\mathbf{x}', \mathbf{y}')) = k(w_{\mathbf{y}}(\mathbf{x}), w_{\mathbf{y}'}(\mathbf{x}')), \tag{7}$$

where $w_{\mathbf{y}}(\mathbf{x})$ refers to the window of $\mathbf{x}$ implied by the label sequence $\mathbf{y}$.

4. Experiments

We assembled a dataset from the Calmodulin Target Database (Yap et al., 2000) consisting of 194 binding sites from 174 proteins. The corresponding protein sequences were acquired from the Swiss-Prot database (Bairoch & Apweiler, 1996). To reduce bias in classifier assessment, proteins with high sequence similarity, 70% or higher, were grouped together such that if one member was selected for training or testing, the rest of the group would be included.

For both the sliding window and structured output classifiers, two amino acid kernels were used for protein windows. The first was the p-spectrum kernel, defined as

$$k_p(w, w') = \sum_{u \in \sum^p} \phi_u^p(w) \phi_u^p(w'), \tag{8}$$

where $\phi_u^p(x) = |\{u | u \in x, \text{len}(u) = p\}|$. This kernel measures similarity between two strings by considering the number of substrings of length p that occur in both sequences (Leslie et al., 2002).

The second kernel we used is based on the physical and chemical properties of amino acids. We extracted 60 properties of amino acids from the Amino Acid Index Database (Kawashima et al., 1999), and represented a window was as a length-60 vector where each feature is the average of that property over the window. The physico-chemical kernel was then used as a Gaussian kernel over this feature space.

For the sliding window classifier, support vector machines (SVMs) were trained using `PyML`, available at `http://pyml.sourceforge.net/`. SVM training data were assembled from windows where the positive samples were the actual binding sites and the negative samples were obtained by selecting windows from the proteins sequences that do not overlap the binding sites. As binding sites account for only a small fraction of amino acids within proteins, model selection over margin penalties and kernel parameters was performed by using stratified cross-validation to account for the data imbalance. Finally, to infer a label sequence, $\hat{\mathbf{y}}$, we selected the window with the largest decision value among a sequence of consecutive windows having the greatest average decision value.

A perceptron algorithm based on (Altun et al., 2003), outlined in Algorithm 1, was used to learn the structured outputs discriminant function. As we are inferring labels of a fixed length for proteins in which the actual binding sites are of various lengths, we only perform a weight update if the distance between the center of the predicted window, $c(\hat{\mathbf{y}})$, and the center of the target $c(\mathbf{y})$, is greater than some radius, ρ. To ensure convergence, a learning rate η was used, where for training epoch i, $\eta = 1/i$.

Algorithm 1 Binding Site Perceptron Training

```
Input: ${(\mathbf{x}_i, \mathbf{y}_i)}_{i=1}^{N}$ {input-output training data}
       $\rho$ {update radius}
       $\eta(epoch)$ {learning rate function}
$epoch = 0$
repeat
  for $i = 1$ to $N$ do
    $epoch += 1$
    $\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} F_\alpha(\mathbf{x}_i, \mathbf{y})$
    if $|c(\hat{\mathbf{y}}) - c(\mathbf{y}_i)| > \rho$ then
      $\alpha_{\mathbf{x}_i,\hat{\mathbf{y}}} \leftarrow \alpha_{\mathbf{x}_i,\hat{\mathbf{y}}} - \eta(epoch)$
      $\alpha_{\mathbf{x}_i,\mathbf{y}} \leftarrow \alpha_{\mathbf{x}_i,\hat{\mathbf{y}}} + \eta(epoch)$
    end if
  end for
until convergence
```

5. Results

The accuracy of our classifiers was assessed at the binding site level. A predicted binding site that overlaps the actual location was considered a correct prediction. In the case where multiple binding sites occur in a single protein, the binding site nearest to the predicted window was considered.

Results of sliding window and structured outputs classifiers that use windows of length 21 are summarized in Table 1. Average accuracies and standard deviations are reported for 20 training-testing splits where for each split, a training dataset is constructed by sampling the input-output pairs without replacement and ensuring sequences with high similarity are included. As a protein of length n has $n-k-1$ labellings, it is important to note that a random classifier has expected accuracy of ≈ 0.05, considering an average binding site length of 21 and an average protein length of 773.

Kernel	Sliding Window	Structured Outputs
2-spectrum	0.49 (0.07)	0.56 (0.05)
Physico-chemical	0.52 (0.07)	0.58 (0.06)

Table 1. Accuracy of calmodulin-binding binding site prediction results for the sliding window and structured outputs classifiers using the p-spectrum kernel and the physico-chemical properties kernel.

The results in Table 1 show that the structured output perceptron outperforms the sliding window SVM. Given the simplicity of the perceptron algorithm it is very encouraging that it performed better than a large margin classifier. This illustrates the advantage of formulating the prediction problem in the structured outputs methodology. A further advantage is not having to perform the usually *ad hoc* step of post-processing the predictions at the amino acid level to infer a predicted label sequence. We expect improved results when we apply the structured SVM to this problem and explore the use of more sophisticated kernels. The issue of multiple binding sites remains to be addressed as well. This can be done either by expanding the output space or iteratively predicting additional binding sites. The methodology developed here can be also applied to other cases of this type of label sequence problem such as identifying disordered regions in proteins and locating binding sites for other target proteins.

References

Altun, Y., Tsochantaridis, I., & Hofmann, T. (2003). Hidden Markov support vector machines. *ICML* (pp. 3–10). AAAI Press.

Bairoch, A., & Apweiler, R. (1996). The SWISS-PROT protein sequence database and its supplement TREMBL. *Nucleic Acids Research*, *24*, 21–25.

Dietterich, T. G. (2002). Machine learning for sequential data: A review. *Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition* (pp. 15–30). London, UK: Springer-Verlag.

Kawashima, S., Ogata, H., & Kanehisa, M. (1999). AAindex: Amino acid index database. *Nucleic Acids Research*, *27*, 368–369.

Leslie, C. S., Eskin, E., & Noble, W. S. (2002). The spectrum kernel: A string kernel for SVM protein classification. *Pacific Symposium on Biocomputing* (pp. 566–575).

Radivojac, P., Vucetic, S., O'Connor, T. R., Uversky, V. N., Obradovic, Z., & Dunker, A. K. (2006). Calmodulin signaling: Analysis and prediction of a disorder-dependent molecular recognition. *Proteins: Structure, Function, and Bioinformatics*, *63*, 398–410.

Vetter, S. W., & Leclerc, E. (2003). Novel aspects of calmodulin target recognition and activation. *European Journal of Biochemistry.*

Yap, K. L., Kim, J., Truong, K., Sherman, M., Yuan, T., & Ikura, M. (2000). Calmodulin target database. *Journal of Structural and Functional Genomics*, *1*, 8–14.