Context Sensitive Vocabulary And its Application in Protein Secondary Structure Prediction

Yan Liu[1], Jaime Carbonell[1], Judith Klein-Seetharaman[1,2], Vanathi Gopalakrishnan[2]
[1]School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-8213, USA
{yanliu, jgc, judith}@cs.cmu.edu

[2]Center for Biomedical Informatics
University of Pittsburgh
Pittsburgh, PA 15213-2582, USA
vanathi@cbmi.upmc.edu

ABSTRACT

Protein secondary structure prediction is an important step towards understanding the relation between protein sequence and structure. However, most current prediction methods use features difficult for biologists to interpret. In this paper, we present a new method that applies information retrieval techniques to solve the problem: we extract a context sensitive biological vocabulary for protein sequences and apply text classification methods to predict protein secondary structure. Experimental results show that our method performs comparably to the state-of-art methods. Furthermore, the context sensitive vocabularies can serve as a useful tool to discover meaningful regular expression patterns for protein structures.

Categories and Subject Descriptors: H.4.4 [Information Systems Applications]: Miscellaneous

General Terms: Algorithms, Experimentation.

Keywords: Biological language, Context sensitive vocabulary, Protein secondary structure prediction.

1. INTRODUCTION

It is widely believed that protein secondary structure prediction can contribute valuable information to discerning how proteins fold in three-dimensions. Given a protein sequence $X_1X_2 \ldots X_N$, where X_i is one of the twenty amino acids, our task is to predict the secondary structure labels for each residue, i.e. $Y_1Y_2 \ldots Y_N$, where Y_i is one of the secondary structures, including helix(H), sheet(E) or coil(C).

The state of art performance for protein secondary structure prediction is achieved by converting it into a classification problem using the window-based methods. The best overall per residue prediction accuracy is 78% on average[2]. However, most prediction methods use features difficult to interpret. Those features fall into two broad categories based on: (a) the use of amino acid sequences alone, such as the propensities of single amino acid for different structures [1] and evolutionary conservation by multiple sequence alignment [7]; and (b) the use of sequence information together with physico-chemical properties of the amino acids, such as hydrophobicity and solvent accessibility. For prediction, various learning algorithms have been applied, such as neural networks [7], linear discriminative analysis (LDA) [6] and support vector machines (SVMs) [5].

To provide insightful analysis for structure formation, we can think of protein sequences being made up of words, then the mapping from protein primary sequence to its structure is intuitively similar to the mapping from documents of human language to grammatical structures and semantic meanings [4]. This motivates our work to introduce a new "context sensitive" vocabulary and then apply text classification methods for human language to protein secondary structure prediction.

2. CONTEXT SENSITIVE BIOLOGICAL VOCABULARY

The protein primary sequence is made up of 20 amino acids and it is still unknown as to what are the most meaningful "vocabulary" for proteins. We use the concept of n-grams, which refers to all amino acid sequences of length up to n within a given sequence (N-grams are similar concepts to the k-mers used by pairwise sequence alignment algorithm, such as BLAST). Initially we used vocabularies consisting of n-grams ($n \leq 5$), but they proved not to be very effective for predictions (for detail see section 3).

To solve this problem, we propose a context sensitive vocabulary. In human language, the same word will have different *semantic meanings* in different contexts. For example, the meaning of the word "bank" in "saving bank" is different from that in "river bank". Similarly, the *physico-chemical properties* of each amino acid will vary in different protein sequences and different positions in the same sequence.

The way we encode the context is through recording the relative positions of the n-grams to the query residue, following the regular expression as "x(-|+)n", where x represents the n-gram of amino acids, -|+ indicates whether the n-gram is before or after the query residue in the protein sequence, and n is the relative position. Table 1 gives an example of the context sensitive biological vocabulary.

For each residue in the protein sequence, we can construct

	$N = 1$	$N = 2$
N-gram	E, C, P, V, N, C, I	EC, CP, PV, VN, NC, CI, IQ
Context Sensitive N-gram	E-3, C-2, P-1, V+0, N+1, C+2, I+2	EC-3, CP-2, PV-1, VN+0, NC+1, CI+2, IQ+3

Table 1: Vocabularies for the 5th residue V in the sequence *PECPVNCIQS* within window size 7

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

a document consisting of the context sensitive words discussed above. Then the assignment of secondary structures is intuitively similar to the classification of document topics in human language. Therefore the state-of-art text classification methods can be applied. In our experiment, we use the SMART *ltc* version of TF-IDF term weighting and Support Vector Machines (SVMs) as classifiers [8].

3. EXPERIMENTAL RESULTS

Two datasets were used for evaluation: one is the RS126 dataset [7, 6], a benchmark that consists of 126 protein sequences with less than 25% similarity, the other is the CB513 dataset with 513 non-homologous protein sequences [2]. The two datasets and profiles of multiple sequence alignments can be downloaded from the web http://barton.ebi.ac.uk/. To evaluate the prediction accuracy, two most common measures are used, i.e. accuracy (Q_3) and segment of overlap (SOV) (for detailed definition see [7, 2]).

Prediction Accuracy We compared our methods with other methods, including PHD [7], DSC [6] and SVM [5] in Table 2. From the results, we can see that (1) the use of context sensitive vocabularies improve the accuracy significantly (by 22% in Q_3) over the common N-gram, which indicates that the context information is very important to protein structures formation; (2) our method can achieve comparable performance with other state-of-art methods.

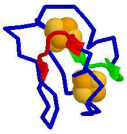
Method	RS126		CB513	
	SOV	Q_3	SOV	Q_3
N-gram vocabulary	57.1	33.0	58.3	32.9
Context sensitive vocabulary	70.7	69.8	71.4	69.6
DSC	69.1	69.0	68.2	68.9
PHD	72.0	70.8	N/A	N/A
SVM	68.7	68.1	70.5	68.0

Table 2: Cross-validation results of different methods on RS126 and CB513 datasets

Discovering Relationships between Sequences and Structures In human language, words with similar concept in *semantics* tend to co-occur frequently in one sentence or one document. Similarly some biological words might have similar *physico-chemical* properties and co-occur in certain structure patterns. Discovery of these relationships might help us understand protein structure formations.

Latent semantic indexing (LSI) is an effective technique in information retrieval to detect word associations [3]. In our experiment, we choose the 2000 largest singular values and compute the word association matrix. The word pairs with highest association scores are {AI+4, CP+0}, {PG-1,GH+0}, {EE-4,EL-3} and the corresponding regular expressions are CPxxAI, PGH, EEL. In order to study whether those results are meaningful, we pick the regular expression CPxxAI as an example and search sequence alignments where the pattern appears. Table 3 shows our findings for the protein ferredoxin (protein data bank ID 1DUR, formerly 1FDX) that shows two occurrences of this pattern. As we can see, CPxxAI describes in both cases the loop region at the C-terminal end of a beta sheet, which are the two enzymes active sites in this protein.

4. CONCLUSION



```
1DUR:. . . PE CP VN CI QGS . . SV CP VG AP NP. . .
MSA1:. . . PE CP VN AI AGD. . GV CP VD AP QP. . .
MSA2:. . . KV CP TG AMNA⊔. . MA CP YD AP QY. . .
MSA3:. . . LR CP MK AI AKE. . PT CP ME AI EL. . .
SS:      . . . HHCCCCCEECC. . HHCCCCCEEE. . .
```

Table 3: Right: Regular expression pattern CPxxAI in protein 1DUR and its multiple sequence alignment (MSA). Left: 3-D structure for protein 1DUR. SS: secondary structure assignment; Red, green: areas matched to the regular expression.

Protein secondary structure prediction is an important step towards understanding structures formation, but most current methods use features difficult to interpret. In this paper, we solve the problem from a language point of view: we introduce the context sensitive biological vocabulary and novelly apply text classification to protein secondary structure predictions. The experimental results show that our method performs as well or better than other methods. More importantly, our context sensitive vocabulary serves as a useful tool to extract meaningful regular expressions for certain structures. We hope our work will lead to the discovery and categorization of new functional motifs within classes of protein structures. Further work involves search for more expressive vocabularies for protein sequences, such as meaningful regular expressions.

5. ACKNOWLEDGMENTS

This work was funded by the National Sceince Foundation (NSF) grant #EIA-0225656.

6. REFERENCES

[1] P. Chou and G. Fasman. Prediction of protein conformation. *Biochemistry*, 13:222–245, 1974.

[2] J. Cuff and G. Barton. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins.*, 34:508–519, 1999.

[3] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[4] M. Ganapathiraju, D. Weisser, J. Seetharaman, R. Rosenfeld, J. Carbonell, and R. Reddy. Comparative n-gram analysis of whole-genome sequences. In *Human Language Technologies Conference (HLT)*, 2002.

[5] S. Hua and Z. Sun. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J Mol Biol.*, 308:397–407, 2001.

[6] R. King and M. Sternberg. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.*, 5:2298–2310, 1996.

[7] B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol.*, 232:584–599, 1993.

[8] Y. Yang and X. Liu. A re-examination of text categorization methods. In *SIGIR'99*, 1999.