Motif Extraction: Normalization of Scores

Y. Fujiwara [1]
yukiko@csl.cl.nec.co.jp

M. Asogawa [2]
asogawa@csl.cl.nec.co.jp

[1] Computer System Research Laboratory, C&C Research Laboratories,
NEC Corporation
4-1-1, Miyazaki, Miyamae-ku, Kawasaki, Kanagawa 216 Japan

[2] Massively Parallel Systems NEC Laboratory, RWCP

Abstract

This paper examines a method to normalize a score of a stochastic motif, represented by a hidden Markov model (HMM). The accuracy of the Z score method, which is one of the score normalization method, is compared with that of the whole search method.

1 Introduction and Methods

The stochastic motif deals with the stochastic nature or the sequence variety resulted from the evolution process, and is represented by an HMM which is commonly used in computational biology[1]. The stochastic motif can calculate a score for a given sequence and compare with the threshold value to predict that the given sequence has the target motif or not[2][3]. The score which is logarithm likelihood, is calculated by multipling the probabilities, therefore a longer sequence tends to have a worse score. The relationship between the length of a sequence (L) and the score (S) is almost linear: $S = AL + B$, where A and B are some constants. Thus, the score must be normalized to compare the scores of different length sequences.

One of the score normalization methods is the Z score method[4]. The Z score method collects the average scores (Ave_L) and the standard deviations (SD_L) of the same length L of the negative examples. Then the method fits $(L, Ave_L + cSD_L)$ to the straight line using the least squares method, where c is some constant determined by the experiments. This means that the fit is weighted by the standard deviation for each length.

To evaluate the accuracy of the Z score method, the whole search method is practiced. The whole search method varies A and B at some intervals and selects the constants which achieves the highest accuracy, which is the average accuracy of positive and negative examples. If the interval is very small, the method achieves the high accuracy, but it takes much time.

2 Data and Results

For experiments, a leucine zipper motif is used. Positive examples, which are the collection of subsequences annotated as leucine zipper (like), were chosen from the Swiss Protein database Release 30. Also, Negative examples were randomly selected, which are not annotated as leucine zipper (like). These positive and examples were randomly divided into three groups. The first group was used for learning HMM topology and parameters. The second group was used for determining A and B. Thus, these two groups were close data. The third group was open data for testing. The HMM topology and parameters were learning for the first group using the iterative duplication method[2][3].

The accuracy of the Z score method is compared with that of the whole search method. Figure 1 shows the accuracy for the second groups and for the third groups. Naturally, the whole method achieves the higher accuracy than the Z score method for the second groups (Figure 1 (a)). However, the constants which achieves highest accuracy for the second groups are not always the best constants of the unknown data, the third groups (Figure 1(b)). This results must come from the constant unstability of the whole search method because the size of the positive examples is small. For the calculation time, the whole method takes almost 50 times as much as the Z score method in the current implementation at intervals of 0.25.

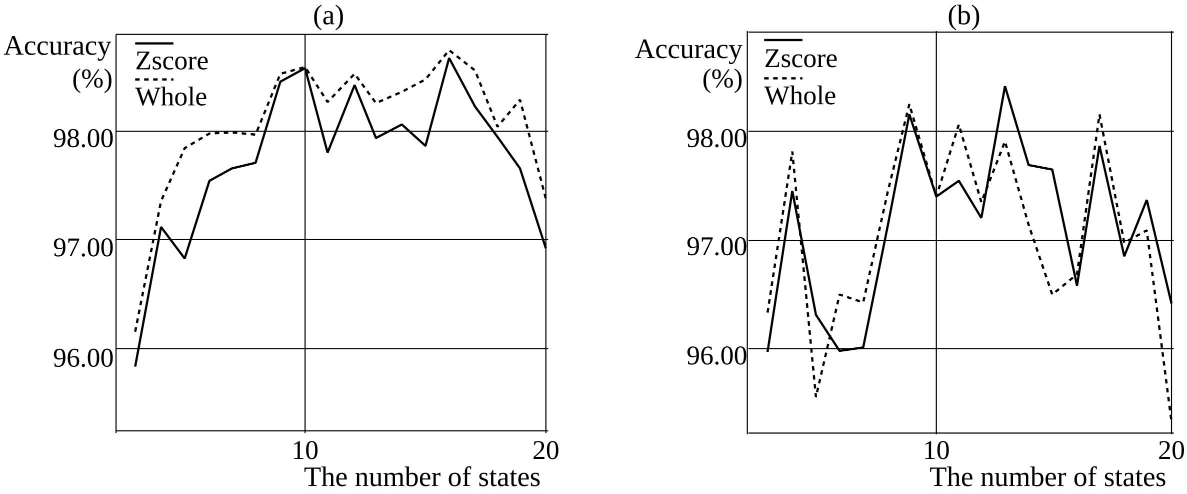

Figure 1: (a) The accuracy for the second group, (b) the accuracy for the third groups

References

[1] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, and D. Haussler, "Hidden Markov models in computational biology: Applications to protein modeling", *J. Mol. Biol.*, Vol. 235, pp. 1501-1531, 1994.

[2] Y. Fujiwara, M. Asogawa, and A. Konagaya, "Stochastic Motif Extraction using Hidden Markov Model", *Proc. 2nd Int. Con. on ISMB*, pp. 121-129, 1994.

[3] Y. Fujiwara, M. Asogawa and A. Konagaya, "Hidden Markov Model to Extract Leucine Zipper Motif", *Proc. Genome Informatics Workshop*, pp. 77-85, 1995.

[4] http:www.caos.kun.nl/GCGdoc/Program_Manual/Multiple_Sequence_Analysis/profilesearch.html