

Classification of C2H2 Zinc Finger Domains Using Support Vector Machines

Takafumi Nagano¹

Nagano.Takafumi@wrc.melco.co.jp

Makiko Suwa²

m-suwa@aist.go.jp

Kiyoshi Asai²

asai-cbrc@aist.go.jp

¹ Advanced Technology R&D Center, Mitsubishi Electric Corp., 8-1-1 Tsukaguchi-Honmachi, Amagasaki City, Hyogo 661-8661, Japan

² Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6 Aomi, Koutou-ku, Tokyo 135-0064, Japan

Keywords: zinc finger, C2H2, sequence analysis, support vector machine, fisher kernel

1 Introduction

Zinc finger proteins include nuclear receptors for steroid hormones and are mainly DNA-binding transcription factors. Thus those are supposed to be target proteins for drug discovery. C2H2 zinc finger gene family is one of the most popular and complex superfamilies. C2H2 zinc finger domains are composed of approximately 25 to 30 amino acid residues including the paired cysteines and histidines that form coordinate bonds with zinc ion. Although C2H2 domains are well-studied, it is difficult to detect the domains with high accuracy by means of homology search or hidden Markov models(HMMs) owing to a wide variety of the sequences.

In this research, we have extended the Support Vector Machine(SVM) based method using the Fisher kernel [1] in order to achieve better accuracy than an HMM. The Fisher kernel extracts a fixed length vector of features known as a Fisher score vector (FSV) from a variable length sequence with an HMM. The method in [1] classifies G-protein coupled receptors (GPCRs) into GPCR subfamilies.

2 Method and Results

The method to discriminate among domains which are detected with little significance by an HMM is proposed. A training data set is constructed from domains detected by an HMM. An SVM is trained to distinguish positive examples from negative examples.

First, the C2H2 domains with positive scores were extracted from protein sequences of SWISS-PROT Release 40.0 using HMMER 2.2 with the profile HMM (zf-C2H2) of Pfam 6.6. The domains whose coordinating residues are neither cysteines nor histidines were removed. The domains which overlapped with domains with uncertain annotation: ATYPICAL, DEFECTIVE, DEGENERATE, INCOMPLETE, POTENTIAL and LOW DNA-BINDING AFFINITY, in the SWISS-PROT database and which didn't wholly correspond with domains in the SWISS-PROT database were removed in order to evaluate the accuracy of the proposed method.

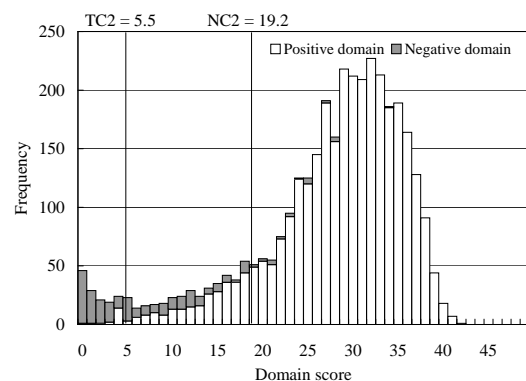


Figure 1: The score histogram of C2H2 domains detected by HMM.

Then the domains which corresponded with domains in the SWISS-PROT database were determined to be positive examples. The others were determined to be negative examples. The score histogram of C2H2 domains detected by the HMM is shown in Fig. 1.

Each Profile HMM in the Pfam database has a trusted cutoff and a noise cutoff: TC2 and NC2, for the domain scores. TC2 is the lowest domain score found in the Pfam full alignment. NC2 is the highest domain score of matches not included in the Pfam full alignment. TC2 and NC2 of zf-C2H2 are 5.5 and 19.2, respectively. Thus we regarded the domains with higher or equal scores to NC2 as positive examples in a training data set and the domains with lower scores than TC2 as negative examples in a training data set. It is noted that the training data set includes misclassifications in order for the proposed method to discriminate among domains detected by an HMM using only a profile HMM, TC2 and NC2. A test data set was composed of the domains with higher or equal scores to TC2 and lower scores than NC2. The training and test data set were made non-redundant. The number of positive and negative examples in the training data set was 2523 and 113, respectively.

All domain sequences were transformed into FSVs on match states in the HMM using nine-component mixture, *uprior.9comp* [1, 3]. As a domain score got lower, the FSV got more scattered from that of the HMM consensus. The number of positive examples in the training data was too large compared with that of negative examples. The positive examples were reduced to 200 domains with the lowest scores in order to capture the sensitive features of the domains around the classification boundary. Each negative example was also refined by replacing some amino acid residues on match states where the viterbi path passing through by those of the positive example which had the minimum mean square distance of FSV. And each negative example was retransformed into FSV. Likelihood ratio scores on delete and insert states where the viterbi path passing through were also calculated and those on the other delete and insert states were set to 0. Because FSVs were based on emission probabilities of match states in the HMM and were not informative enough for an SVM to be trained to distinguish positive examples from negative examples in the test data set with high accuracy.

Then we trained a linear ν -SVM[2] using FSVs and the likelihood ratio scores on delete and insert states. The ν -SVM has the advantage of using a parameter ν on controlling the number of margin errors and support vectors. The positive accuracy (TP / (TP + FN)) and negative accuracy (TN / (TN + FP)) on the test data set when trained with $\nu = 0.1$ were 75.0% and 76.9%, respectively. The positive and negative accuracy of the HMM were 73.0% and 72.5%, respectively. The performance of the HMM were evaluated as follows. Changing a threshold from TC2 to NC2, domains with higher or equal scores to the threshold were considered to be classified as positive examples and domains with lower scores than the threshold were considered to be classified as negative examples. Then the result at the threshold where both positive and negative accuracy were high was regarded as the performance of the HMM.

The proposed method showed better performance than the HMM. We note that the proposed method should be applicable to a variety of domains, since it didn't make use of specific characteristic of C2H2 domains.

References

- [1] Rachel, K., Kevin, K., and David, H., Classifying G-protein coupled receptors with support vector machines, *Bioinformatics*, 18(1):147–159, 2002.
- [2] Schölkopf, B., Smola, A.J., Williamson, R.C., and Bartlett, P.L., New support vector algorithms, *Neural Computation*, 12:1207–1245, 2000.
- [3] Sjölander, K., Karplus, K., Brown, M.P., Hughey, R., Krogh, A., Mian, I.S., and Haussler, D., Dirichlet mixtures: A method for improving detection of weak but significant protein sequence homology, *CABIOS*, 12:327–345, 1996.