Motif Extraction using an Improved Iterative Duplication Method for HMM Topology Learning

Yukiko Fujiwara[1], Minoru Asogawa[2], Akihiko Konagaya[2]

[1]Computer System Research Lab., C&C Research Labs., NEC Corporation

[2]Massively Parallel Systems NEC Laboratory, RWCP*

[1,2]4-1-1, Miyazaki, Miyamaeku, Kawasaki, Kanagawa 216, Japan

yukiko@csl.cl.nec.co.jp, asogawa@csl.cl.nec.co.jp, konagaya@csl.cl.nec.co.jp

In this paper, Hidden Markov Models (HMMs) are utilized to represent motifs. Motifs are the preserved sites in the evolution process and are considered to represent the function or structure of proteins. It is compulsory to use direct periodic topology and the OR rule to represent motifs. The direct periodic topology is necessary to represent cyclic structures, such as global loops for α helices. The OR rule is necessary to represent different families or types in motifs. Therefore, to represent motifs, HMMs must contain direct periodic topology and OR rule.

In order to obtain an optimal HMM topology for a motif, we improved the "*iterative duplication method*"[1]. In this method, a small fully-connected HMM is gradually expanded by a state splitting. The initial number of the states is set to three in the current implementation. It iterates through a parameter estimation phase followed by a topology modification phase. It terminates when the accuracy of model size selection data is maximum.

For parameter estimation, the Baum-Welch algorithm is used on several initial HMMs with randomized parameters to avoid converging in the local maximum. Then, the best HMM is chosen based on the likelihood of training data.

In the topology modification phase, negligible transition deletion, negligible state deletion, splitting state selection and state splitting is performed. The negligible transition deletion is performed by deleting the transitions with negligible transitional probability, the threshold is $\epsilon = \max(\epsilon_1, r)$, where ϵ_1 is a smoothing value and r is a convergence radius. The purpose of deleting the transitions is to keep the topology space small and consequently to reduce the computation for the parameter estimation. The negligible state deletion is performed by deleting states with negligible initial probabilities and negligible incoming transitions. This state deletion also reduces the training computation and has a good effect for calculating the criteria of splitting state. The

*RWCP: Real World Computing Partnership

splitting state selection is performed based on the transition ambiguity and the observation ambiguity described below. The transition ambiguity of the state S_j is defined as $in(j) + out(j)$. Where the $in(j)$ represents the sum of the probabilities of incoming transitions and the $out(j)$ represents the sum of the weighted probabilities of outgoing transitions. The observation ambiguity is calculated as follows. First, for each state S_i, the observation distribution on the condition that the previous observation is v_k is estimated. These distributions are calculated using the forward-backward algorithm. In the case of amino acid sequences, twenty distributions are calculated for each state. Next, these distributions at each state S_i are divided into two groups. Then, for each state, the distribution difference of the two groups is calculated by squared error. Similarly, for each state S_i, the observation distribution on the condition that the next observation is v_k is estimated. After that, the observation ambiguity for each state is defined the sum of the two distribution difference.Finally, the sum of the regularized transition ambiguity and observation ambiguity is calculated for each state. Then, the state with the maximum value is chosen for the splitting state. This criteria of the splitting state selection is the main contribution of this paper. The criteria of the previous method relied only on the transition ambiguity, such as the number of transitions. This criteria is considered more details of the transitions and the expected observation differences by the splitting. The state splitting is performed as follows. For selected splitting state S_c, a copy of it (denoted as S_{new}) is made so that S_{new} has the same transitions as S_c. If the state S_c has a self-loop, S_{new} has a self-loop and the transitions from S_c to S_{new} and from S_{new} to S_c.

Using this method, we obtain an HMM for a leucine zipper motif. Positive data are the subsequences annotated as leucine zipper (like) in the Swiss-Prot database Release 30. Negative data are the collection of subsequences which are not annotated as leucine zipper (like). There are 183 positive data and 377429974 negative data. In order to evaluate the prediction performance, positive or negative data is divided into 3 groups. The first group is used for learning HMM topology and parameters. The second group is used for selecting the model size and the threshold values. Thus, these two groups are closed data. The third is open data for testing. The accuracy of the improved method compares with that of the previous method. The prediction accuracy of the improved method is 96.48 percent for open data 97.87 percent for positive data and 95.09 percent for negative data. The accuracy of the previous method was 95.85 percent for open data: 95.74 percent for positive data and 95.95 percent for negative data. The obtained HMM of the improved method contains global loops for α helices in the leucine zipper motif.

References

[1] Y. Fujiwara, M. Asogawa and A. Konagaya: "Stochastic Motif Extraction using Hidden Markov Model", p121-129, Proceedings Intelligent Systems for Molecular Biology (1994).