Stochastic Motif Extraction Using Hidden Markov Model

Yukiko Fujiwara Minoru Asogawa Akihiko Konagaya
Massively Parallel Systems NEC Laboratory, RWCP*
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

Abstract

In this paper, we study the application of an HMM (hidden Markov model) to the problem of representing protein sequences by a stochastic motif. A stochastic protein motif represents the small segments of protein sequences that have a certain function or structure. The stochastic motif, represented by an HMM, has conditional probabilities to deal with the stochastic nature of the motif. This HMM directly reflects the characteristics of the motif, such as a protein periodical structure or grouping. In order to obtain the optimal HMM, we developed the *"iterative duplication method"* for HMM topology learning. It starts from a small fully-connected network and iterates the network generation and parameter optimization until it achieves sufficient discrimination accuracy. Using this method, we obtained an HMM for a leucine zipper motif. Compared to the accuracy of a symbolic pattern representation with accuracy of 14.8 percent, an HMM achieved 79.3 percent in prediction. Additionally, the method can obtain an HMM for various types of zinc finger motifs, and it might separate the mixed data. We demonstrated that this approach is applicable to the validation of the protein databases; a constructed HMM has indicated that one protein sequence annotated as "leucine-zipper like sequence" in the database is quite different from other leucine-zipper sequences in terms of likelihood, and we found this discrimination is plausible.

Keywords: Hidden Markov Model (HMM), motif extraction, HMM topology learning, iterative duplication method, database validation

Introduction

Extracting a motif from protein sequences is an important problem. Motifs, the preserved sites in the evolution process, are considered to represent the function or structure of the proteins. This motif extraction problem increases in importance as many protein sequences are revealed, because the rate of sequencing far exceeds that of understanding the structures.

*RWCP: Real World Computing Partnership

Until now, a symbolic pattern was used to represent a motif. For example, the pattern of the leucine zipper motif, a well-known motif for the DNA binding proteins, is "L-X(6)-L-X(6)-L-X(6)-L-X(6)-L" representing a repetition of Leucine with any six residues. One of the issues in motif representation is the exception handling caused by the variety of amino acid sequences. Konagaya(Konagaya & Kondou 1993) employed a stochastic decision predicate, which consists the conjunctive and disjunctive patterns and their probability parameter to represent the exceptions in the motif.

However, using a pattern representation can not produce satisfactory classification accuracy. For example, the accuracy of leucine zipper motifs is only 14.8 percent. This is because proteins usually have various sequences corresponding to different species, even around motifs. In leucine zipper motifs, the repeated L's (Leu) tend to change to other amino acids, such as V (Val), A (Ala), M (Met). Such variations are considered to be related to the evolution process of organisms. Thus, these variations might be some systematical relationships, i.e., the variations of amino acids at a residue relate to the neighbor residues. These systematical relationships represent biological characteristics. An HMM can represent these systematical relationships or biological characteristics. Therefore, in this paper, a stochastic motif using an HMM is employed to achieve high classification accuracy.

It is desirable to extract these biological characteristics from training data only. They must be reflected in the HMM topology. For example, when protein structures of training data are periodical such as helices, it is expected that the obtained HMM topology is periodical. When training sequences comes from two different subgroups or families, it is expected that the obtained HMM topology branches in two. For this purpose, general HMMs containing global loops are needed instead of left-to-right models commonly used in speech recognition. Accordingly, determining the HMM topology is one of the problems to solve, because there are lots of candidate topology in general HMMs.

One of the methods to determine the HMM topol-

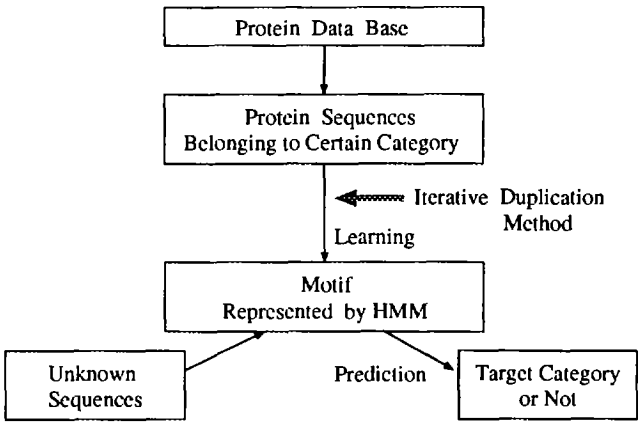

Figure 1: Outline of motif extraction

ogy is to train from a large fully-connected HMM and delete negligible transitions. However, the HMM resulting from a fully-connected one may be very complex and difficult to interpret. Moreover, it takes quite a bit training time in order to optimize numerous free parameters.

Thus, a new method, "*the iterative duplication method*", is developed for HMM learning(Fujiwara & Konagaya 1993). The method enables us to obtain an optimal HMM topology for the given training sequences, as well as optimal HMM parameters for the network. It starts from a small fully-connected network and iterates the network generation and parameter optimization. The network generation prunes transitions and adds a state according to the previous topology. This method obtains simpler HMM topology in less time than the one obtained from a fully-connected model. As a result of this method, for example, the accuracy of leucine zipper motifs is 79.3 percent. It is high when compared with accuracy of 14.8 percent when using the symbolic representation. Figure 1 shows the outline of this motif extraction.

The validation of protein databases using HMMs is also discussed. One of the HMMs we have constructed has indicated that a protein sequence annotated as "leucine-zipper like" in the database is quite different from other leucine zipper sequences in terms of likelihood, and we found this annotation is questionable. Additionally, our method also revealed that some sequences without motif annotations should be discriminated as leucine zipper motifs.

This paper is organized as followings. First, we explain HMMs and related works on HMMs. Then, we explain the "*iterative duplication method*" for HMM learning and the predicting method. After that, we give the experimental results of two motif extractions and discuss examples of validating the database. Finally, we discuss the *iterative duplication method* in more detail.

HMMs

Overview

An HMM is a nondeterministic finite state automaton that represents a Markov process. HMMs are commonly used in speech recognition(Nakagawa 1988), and recently have been applied to protein structure grammar estimation(Asai, Hayamizu, & Onizuka 1993) and protein modeling(Haussler *et al.* 1993), (Baldi *et al.* 1994).

An HMM is characterized by a network with a finite number of parameters and states (see Figure 2). Parameters represent initial probabilities, transition probabilities, and observation probabilities. At discrete instants of time, the process is assumed to be in one state and an observation (or output symbol) is generated by the observation probability corresponding to the current state. This state then changes depending upon its transition probability.

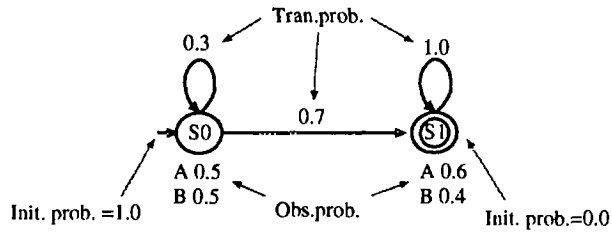

Figure 2: An example of HMM (left-to-right)

A special type of HMM, called a left-to-right model in Figure 2, is commonly used in the case of speech recognition. In this model, states are linearly connected with self-loop transitions; a state visited once is never revisited at a later time. This is because there is little requirement to deal with periodic structures in speech recognition. However, such periodic structures are rather common in amino acid sequences and have great significance for constructing a geometric structure. Therefore, we adopt a general HMM containing global loops.

The correspondence between motifs and HMMs is the following. The training set is the portions of amino acid sequences that have the same structure or function. An HMM is expected to model the training proteins in terms of discrimination. The alphabet used for the output symbols corresponds to 20 amino acids. The test sequence is the portion of an amino acid sequence which might have the target structure or function. The result is the likelihood of the test sequence calculated by tracing all possible transition paths that observe the test sequence in the HMM.

To use a trained HMM as a classifier, we first define a threshold value according to the probabilities generating positive examples. The probability generated by a given sequence is compared against the threshold

Table 1: Zinc Finger in Prosite R25

Type	pos	neg	Consensus patterns in Prosite
C2H2	836	48	C-x(2,4)-C-x(12)-H-x(3,5)-H
C4	315	8471	C-x(2,7)-C-x(0,28)-C-x(1,4)-C
C3H	33	152	C-x(2)-C-x(4)-H-x(4)-C or C-x(2)-C-x(12,13)-H-x(4)-C
C3HC4	30	0	C-x(2)-C-x(0,42)-C-x-H-x-[LIVMFY]-C-x(2)-C-[LIVMA]-C-x(0,42)-C-x(2)-C
GATA	30	0	C-x-N-C-x(4)-T-x-L-W-R-R-x(3)-G-x(3)-C-N-A-C
C8C5C3H	8	5	C-x(8)-C-x(5)-C-x(3)-H
PARP	8	0	CK-x-C-x-[QE]-x(3)-K-x(3)-R-x(16,18)-W-[HY]-H-x(2)-C
Others	65	–	
Total	1327	8676	

value, and the sequences producing larger values are classified as the motif. One of the great advantages of using HMMs is that we can quantify similarity between the test sequence and the training set by comparing their likelihood on the HMM.

Related Works

Haussler et.all(Haussler *et al.* 1993) use HMMs for stochastical modeling and multiple alignment of globins. Baldi et.al.(Baldi *et al.* 1994), (Baldi & Chauvin 1994) use HMMs for globins, immunoglobins and kinases. Since they chose one of the left-to-right model, they cannot treat global loops except self-loops. Asai et.all(Asai, Hayamizu, & Onizuka 1993) use HMMs for secondary structure prediction. They propose a method to construct a large HMM from smaller HMMs using a protein structure grammar. However, the grammar depends on human knowledge at the current stage. As for automatic learning of HMM network topology, Takami(Takami & Sagayama 1991) proposes a state splitting method for speech recognition. It starts from a small HMM and increases the number of states by choosing better state splitting; although the model is restricted to left-to-right models. Our approach is more general and can deal with any network topology.

Motif Extraction using HMM

Training Data and Test Data

For the purpose of extracting a leucine zipper motif, 112 positive examples, which are the collection of subsequences annotated as leucine zipper (like), were chosen from the Swiss Protein database Release 22. Also, 112 negative examples were randomly selected, which satisfy the Prosite (motif database) pattern "L-X(6)-L-X(6)-L-X(6)-L-X(6)-L", a repetition of leucine and any six residues(Aitken 1990). Naturally, those negative examples are similar to positive examples in terms of symbolic representation. Randomly selected, 80 percent of the positive subsequences are used for training, and the remaining positive and all negative examples are used for prediction performance evaluation.

Additionally, a zinc finger motif, describing a nucleic acid-binding protein structure, was also extracted with our method. There are 1327 subsequences annotated as zinc fingers in the Swiss Protein database Release 25. The training data were chosen to be 90 percent of these positive data. The test data is the rest of the positive data and the 8676 negative data containing the motif pattern (See Table 1).

Learning Algorithm

input: (protein) sequences and
a small fully-connected HMM.
initialization:
optimize parameter for the HMM.
choose the best HMM on likelihood
as a seed.
repeat
network generation:
delete negligible transitions.
copy the most connected state.
parameter optimization:
optimize parameters for the new topology.
choose the best HMM on likelihood.
until sufficient accuracy is obtained.
output: the resulting HMM.

Figure 3: Iterative Duplication Method

In order to obtain the optimal HMM topology for the given training sequences, an "*iterative duplication method*(Fujiwara & Konagaya 1993)" is used. This method also produces the optimal HMM parameters for the network. The method includes transition network generation and parameter optimization. The method is summarized in Figure 3. It starts from a small fully-connected network. In order to avoid converging in the local maximum, many initial HMMs with random parameters are prepared. The Baum-Welch algorithm is used for parameter optimization. Network generation is performed by copying one node

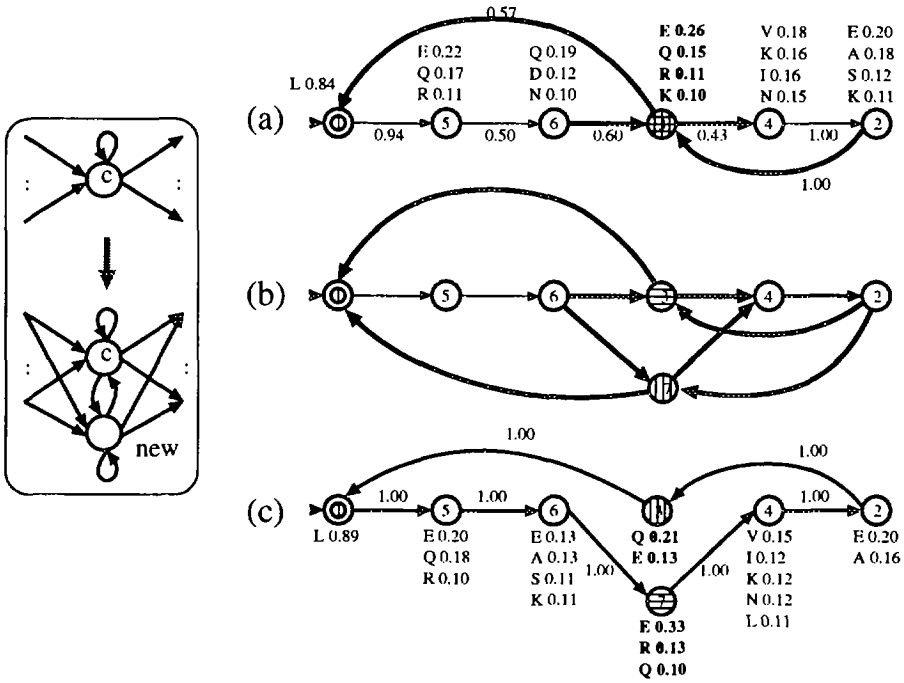

Figure 4: A step of learning. (Left) A general rule for a duplication. (Right) An example of the step from six to seven. (a) The resulting HMM with 6 states after parameter optimization and negligible transition deletion. (b) A new network by a hatched state copy. (c) An obtained HMM with 7 states after parameter optimization and negligible transition deletion.

selected from the current network. The method iterates the network generation and parameter optimization phases until sufficient discrimination accuracy is obtained.

The details of network generation follow. First, we delete the transitions with negligible transitional probability, that is less than $\epsilon = \max(\epsilon_1, r)$, where ϵ_1 is a smoothing value and r is a convergence radius. Next, for each state S_i except the final state, we count the number of incoming and outgoing transitions of the state, that is the number of transitions from the state S_i plus that of transitions to the state S_i. Then we select the state with the largest number (denoted as S_c) and make a copy of it (denoted as S_{new}) so that S_{new} has the same transition with 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 (see Figure 4).

The purpose of deleting the negligible transitions is to restrict the network topology space and eventually to reduce the training cost for parameter optimization. The reason to split the most connected node is that it might represent overlapping of independent states. In this case, the network topology may become simpler by splitting the states.

Figure 4 shows an example of such a case[1]. In Figure 4 (a), the most connected state is a hatched state which outputs E (Glu) with probability 0.26, Q (Gln) with probability 0.15 and so on. By splitting the state into two states, we will obtain a new network which has additional transitions represented by bold lines (see Figure 4 (b)). However, the network can be become simpler, if the most transitions become negligible after parameter optimization (see Figure 4 (c)).

In each step, this algorithm products an optimal HMM for the training data with each number of states. Selecting the HMM with highest prediction accuracy, we obtain the optimal number of states for the given data.

Predicting Method

Prediction is performed by comparing likelihood, that is the probability of generating a given sequence and a threshold value obtained from training data. If a sequence achieves higher likelihood than the threshold value, then it is predicted to have the target motif, that is, the same structure and/or function.

In the current implementation, the threshold value is represented by the worst observed probability in the training set with the same length as a given sequence. Such threshold dependency to length is regarded as based on the fact that longer sequences tend to have worse observation probabilities. More careful study is needed to discuss more meaningful thresholds. One solution would be to use discrimination analysis with data of various lengths (see Figure 5). This discrimination analysis is used for the zinc finger motif analysis.

The total prediction performance is measured by the following equation,

$$(accuracy) = 1.0 - \frac{1}{2} \times \left(\frac{E^+}{N^+} + \frac{E^-}{N^-}\right),$$

[1]The transitions unrelated to our explanation are omitted in Figure 4.

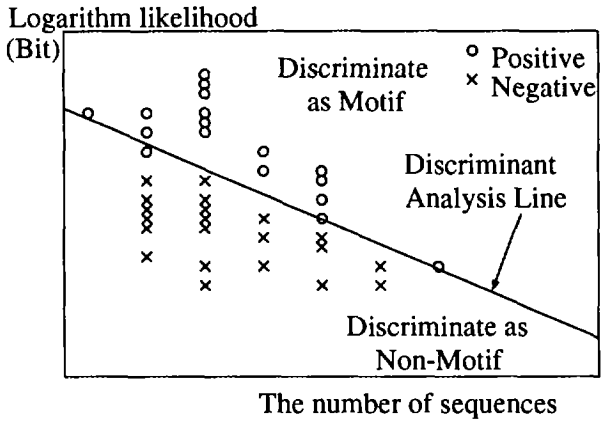

Figure 5: Discrimination Analysis

where N^+ is the number of positive examples, N^- is the number of negative examples and E^+ is the number of errors occurred in the positive examples, that is, the number of the positive data categorizing as the negative by the HMM. E^- is the number of negative errors.

Evaluation

Experimental Results

Table 2 shows the result of cross-validation for leucine zipper motifs. Positive data is divided into 5 groups and tested with both negative and positive data. The average prediction accuracy is 99.1 percent for training data and 79.3 percent for test data; 81.3 percent for positive data and 77.3 percent for negative data. Note that in case of the symbolic motif representation as in Prosite, the average prediction accuracy for test data is just 14.8 percent; 29.5 percent for the positive data and 0.0 percent for the negative data. The low prediction performance of symbolic representation mainly results from the biasing in the negative test set which is chosen from the sequences similar to the positive set. It is apparent HMMs are more powerful than symbolic representation in terms of prediction performance.

Table 2: Prediction accuracy (leucine zipper)

	training	test		
	pos.data	pos.data	neg.data	average
test0	98.9%	81.8%	83.0%	82.4%
test1	100.0%	91.3%	65.2%	78.2%
test2	98.9%	68.2%	83.9%	76.1%
test3	98.9%	87.0%	78.6%	82.8%
test4	98.9%	77.3%	75.9%	76.6%
Ave.	99.1%	81.3%	77.3%	79.3%

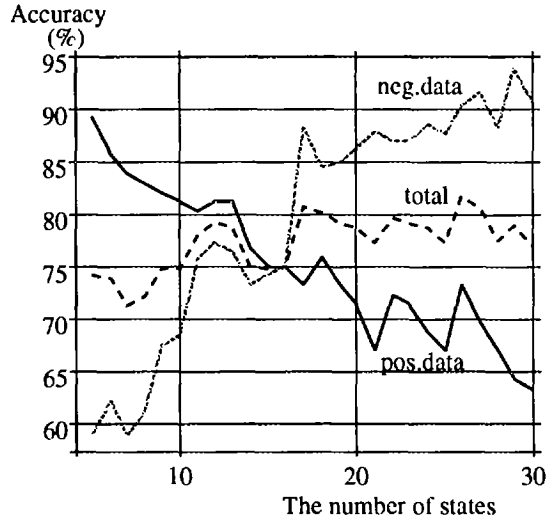

Figure 6: Prediction performance (leucine zipper)

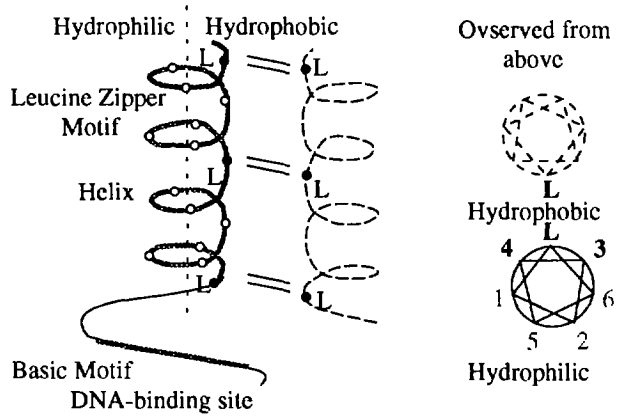

Figure 7: (Left) Biological structure of a leucine zipper motif. (Right)The helical wheel.

Figure 6 shows the test data prediction performance with respect to the number of states for the leucine zipper motifs. It shows that as the number of states increases, the prediction performance for negative data increases, but for positive data, it decreases. This is natural since the expressive power of HMMs increases as the number of states, and may over-fit the training data. Criterion such as the MDL principle may be helpful for avoiding the over-fitting but further study is required. In the current implementation, HMMs with twelve states are chosen, because we regard positive accuracy to be more important than negative.

Figure 8 shows an HMM for leucine zipper motifs obtained using the *iterative duplication method*. This HMM contains global loops corresponding to the "helical structure" in the leucine zipper motif. Such helical

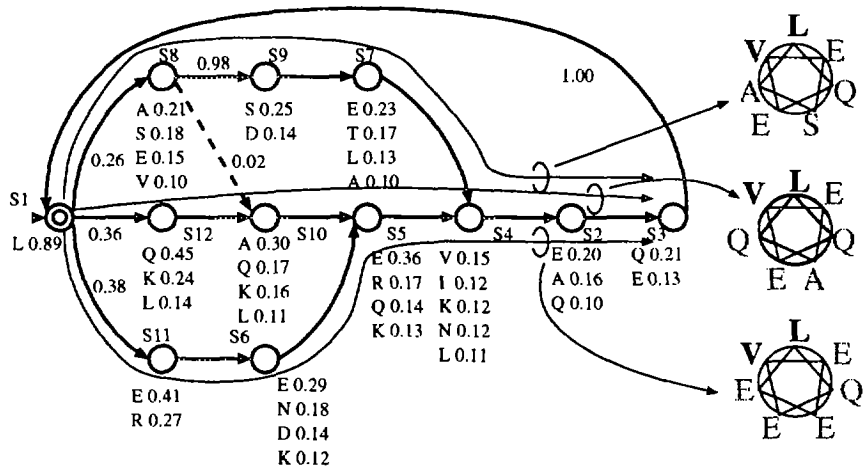

Figure 8: (Left) An HMM (leucine zipper). (Right) The helical wheel, i,e. , helices observed from the head at HMM paths. Hydrophobic, hydrophilic and the other amino acids are described by bold, pale and broken letters, respectively. The characters on the circles are the most frequently observed amino acids at each state.

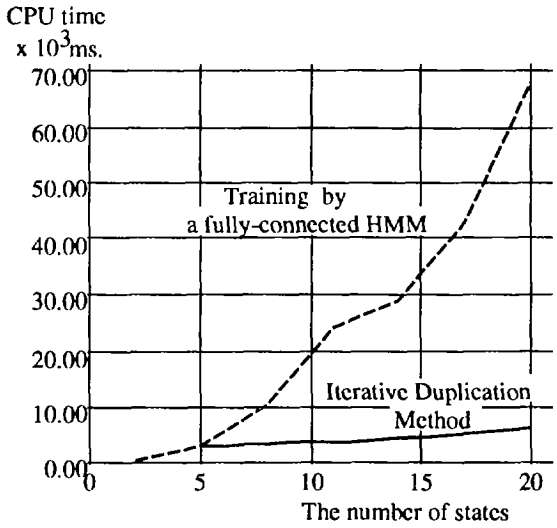

Figure 9: Learning time (leucine zipper)

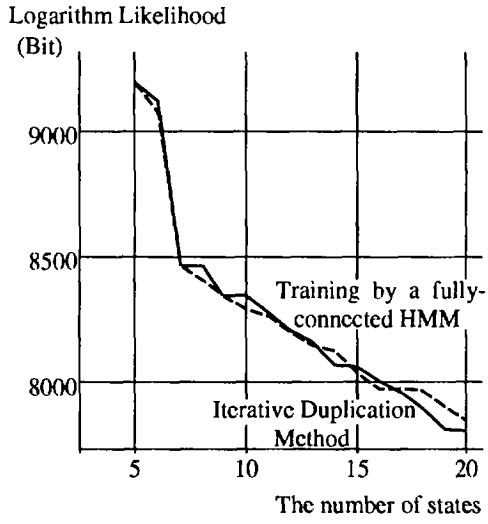

Figure 10: Likelihood (leucine zipper)

structure shown in the figure 7 is caused by the existence of seven amino acids per each two periods. This is because a pair of aligned leucines forms a zipper-like structure. On the sight of Figure 7, this characteristic can be seen in the circles, the helices viewed from above. These circles show that there are many hydrophobic amino acids on one side around combined leucines and many hydrophilic amino acids on the other side. This tendency of hydrophilic and hydrophobic amino acids is one of the characteristics of the helices, and is called a helical wheel. As compared with this, in the figure 8, each circle at the right corresponding to each HMM pass has a similar characteristic to helical wheel. The characters on the circles are the most frequently observed amino acid in each state. In order to see the helical wheel, hydrophobic amino acids, such as I (Ile), V (Val), L (Leu), F (Phe), C (Cys) are described by bold letters in the following. On the other hand, hydrophilic amino acids, such as R (Arg), K (Lys), N (Asn), D (Asp), Q (Gln), E (Glu), H (His) are described by pale letters. Another M (Met), A (Ala), G (Gly), T (Thr), S (Ser), W (Trp), Y (Tyr), P (Pro) are described by broken letters. These circles show three kinds of helical wheels. Therefore, the *iterative duplication method* automatically extracted the helical structures and characteristics from the positive data.

A similar result can be obtained when using a large fully-connected HMM. However, it takes much longer time than our method (see Figure 9, 10).

Table 3 shows the result of cross-validation for zinc finger motifs. Positive data is divided into 10 groups

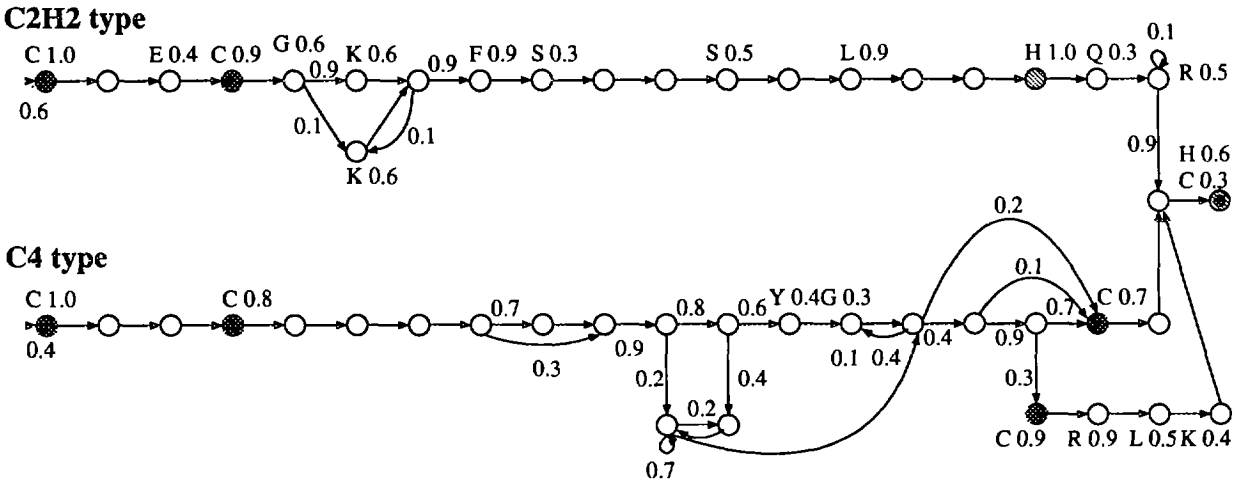

Figure 12: An HMM (zinc finger)

Table 3: Prediction accuracy (zinc finger)

	training	test		
	pos.data	pos.data	neg.data	average
test0	93.4%	83.2%	87.8%	85.5%
test1	93.7%	95.5%	80.8%	88.1%
test2	93.1%	90.2%	78.1%	84.2%
test3	92.3%	93.3%	80.9%	87.1%
test4	92.6%	88.8%	77.5%	83.2%
test5	91.5%	83.6%	80.1%	81.8%
test6	93.8%	91.8%	81.9%	86.8%
test7	93.1%	89.4%	81.9%	85.6%
test8	92.3%	86.4%	84.2%	85.3%
test9	94.1%	91.6%	84.2%	87.9%
Ave.	93.0%	89.4%	81.7%	85.6%

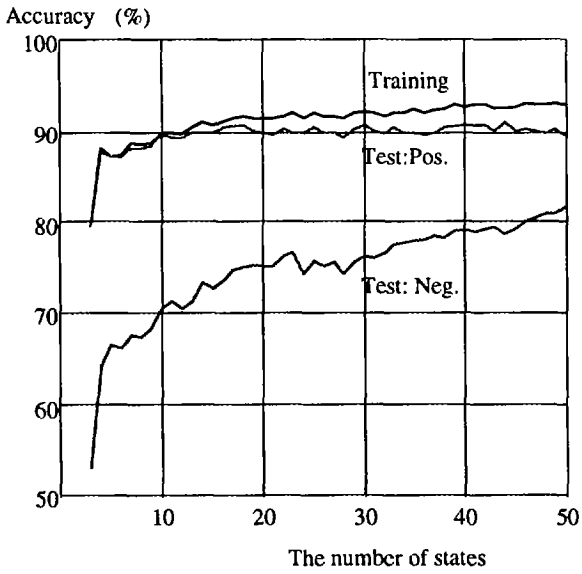

Figure 11: Prediction performance (zinc finger)

and tested with both negative and positive data. Figure 11 shows the prediction performance with respect to the number of states for the zinc finger motifs. The average prediction accuracy is 93.0 percent for training data and 85.6 percent for test data; 89.4 percent for positive data and 81.7 percent for negative data. Using the symbolic representation described in Table 1, the accuracy is 47.5 percent (95.0 percent for positive data and 0.0 percent for negative data). These leucine zipper and zinc finger motifs are represented by ambiguous symbolic patterns, i.e., they are weak motifs.

Figure 12 is an HMM for a zinc finger motif. The biological structure of the zinc finger is shown in figure 13. Mixed data, e.g., C2H2 and C4 type, are used for training; however these mixed data might be separated into several passes on the basis of their types by our method.

The Prosite gives more details, i.e., "generally, but not always, the residue in position +4 after the second cysteine is an aromatic residue, and that in position +10 is a leucine" in case of C2H2 type. Our method reveals these tendencies. Moreover, it reveals the tendencies in the other position (See in Figure 12).

Validation of Database

Since the HMM gives a quantitative value with respect to the similarity of a sequence to the training

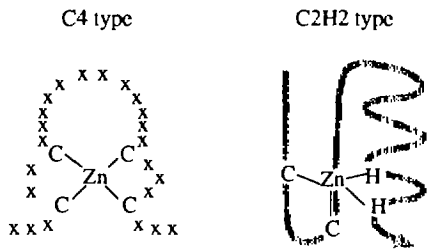

Figure 13: Biological structure (zinc finger)

sequences, it is also applicable for the database validation. For example, in case of a leucine zipper motif, the likelihood of positive and negative examples with length 29 is shown in Figure 14. As shown in the figure, one data (KU7_HUMAN) which has an annotation "leucine zipper motif like" is far from the other positive examples. Therefore, this annotation of KU7_HUMAN is questionable. According to the original paper mentioning KU7_HUMAN, the authors stated that they are not fully confident KU7_HUMAN is the leucine zipper motif. Note that KU7_HUMAN is included in the training data.

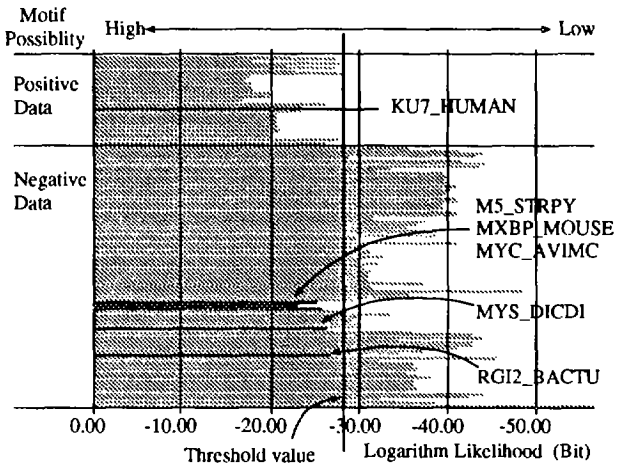

Figure 14: Each likelihood (leucine zipper)

At the same time, the HMM reveals that MYC_AVIMC has a subsequence that achieves high likelihood. However, it has no annotation. MYCs in other species, MYC_FELCA, MYC_HUMAN, MYC_MOUSE have annotations about "leucine zipper motif". There are five such data with no annotations for motifs. These indistinct data were also not omitted in calculating accuracy.

The questionable cases are shown in Figure 15. The circles show helices observed from above. Data are described from the inside out. ATF6_HUMAN is one of

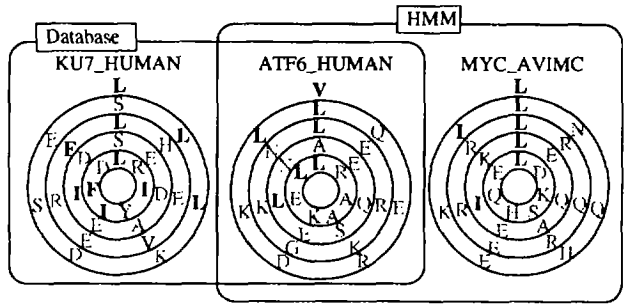

Figure 15: An example of database Validation

the positive data. Compared to this, KU7_HUMAN breaks the helical wheel with respect to the distribution tendency of hydrophilic and hydrophobic amino acids. MYC_AVIMC not annotated motif has a typical leucine zipper motif, aligned leucines and the helical wheel.

These results show that HMMs have good potential for the database validation.

Discussion

In order to represent a motif, an HMM has some advantages over a symbolic pattern. It deals with the stochastic nature of the motif. For example, a leucine zipper motif forms a helical structure, α helix, characterized "helical wheel" that tends to partition hydrophilic and hydrophobic residues. This tendency is difficult to represent accurately by a symbolic pattern, so the symbolic pattern "L-X(6)-L-X(6)-L-X(6)-L-X(6)-L" is far from being a specific pattern. On the other hand, the HMM represents the tendency, enabling us to obtain high accuracy in prediction. Moreover, the HMM represents the relationship among amino acids. For example, figure 8 in the previous section shows that if G (Gly) is next to L (Leu), the amino acid next to G tends to be A (Ala), Q (Glu) or K (Lys).

The *iterative duplication method* seems to produce the optimal HMM topology using only training data. This method gradually grows the HMM topology containing global loops.

The method initially chooses a small HMM topology for training data after parameter optimization and negligible transition deletion. In this topology, each state might contain the union of independent characteristics, e.g., output probabilities and state connections, since the number of states is insufficient.

Then, the method tries to find an overlapping state for dividing its characteristics. The most connected state is selected for division, because this state has many transitions from or to various states. In the current implementation, we randomly select one of the most connected states. Some selection criteria may be studied in the future.

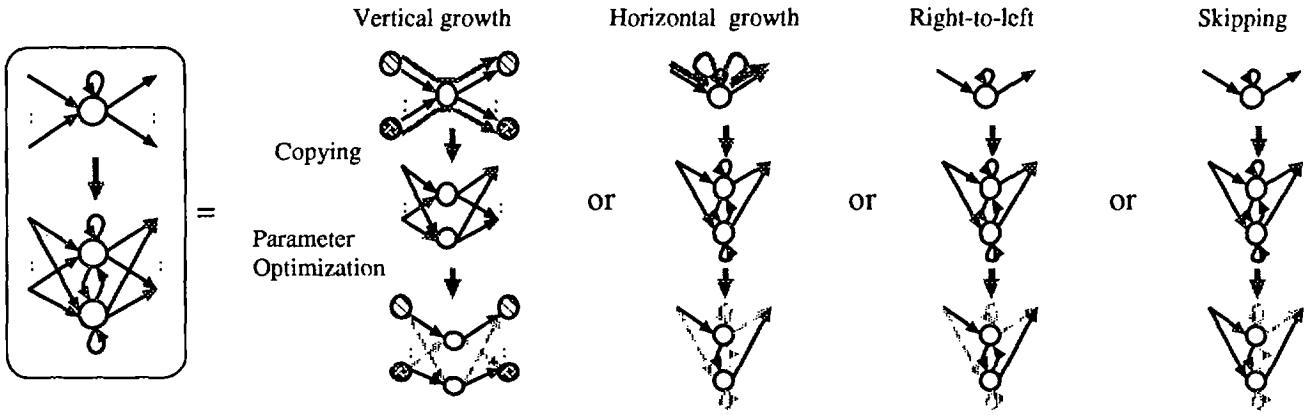

Figure 16: HMM topology growth

Copying a state might grow the HMM topology. This is clear after parameter optimization at the next step (see Figure 16). When two branches are needed at the state, the HMM grows in the vertical direction. On the other hand, when it is necessary to lengthen at a state, it grows in the horizontal direction. Moreover, it grows more complex structure, such as right-to-left transitions, or skipping transitions over states, if needed.

In each step, trained parameters might be used as the next initial parameters. However, these parameters tend to lead a local optimal topology, because the Baum-Welch algorithm converges to the local maximum likelihood. Moreover, this iterative method expands the error. Therefore, random initial parameters are used in our method. Stochastic search algorithm such as a genetic algorithm, would be applicable in stead of random search.

Conclusion

An HMM is capable of representing a stochastic motif well. According to the experience of leucine zipper motif extraction, the HMM shows higher discrimination performance than symbolic motif representation. The HMM is also useful for the validation of annotated comments in amino acid sequence database. In fact, one ambiguous entry was detected by the generated HMM. Additionally, some data is found that they omit annotations of leucine zipper motifs. As for the learning performance, the *iterative duplication method*, increasing the number of states step by step, produces the optimal HMM containing global loops from the training data only. It greatly reduces convergence speed compared to Baum-Welch algorithm for fully connected HMMs.

Acknowledgements — The authors would like to thank Dr. Shimada, the director of Real World Computing Partnership, Dr. Yamamoto, the director of C&C Research Laboratories, NEC Corporation, Dr. Koike, the senior manager of Computer System Laboratory, NEC Corporation and Mr. Kajihara, the assistant manager of massively parallel systems NEC laboratory.

References

Aitken, A. 1990. *Identification of Protein Consensus Sequences.* Ellis Horwood Limited.

Asai, K.; Hayamizu, S.; and Onizuka, K. 1993. hmm with protein structure grammar. *Proceedings of the Twenty-sixth Hawaii International Conference on System Science* I:pp783–791.

Baldi, P., and Chauvin, Y. 1994. smooth on-line learning algorithms for hidden markov models. *Neural Computation* 6(2):pp307–318.

Baldi, P.; Chauvin, Y.; Hunkapiller, T.; and McClure, M. A. 1994. hidden markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences of the USA* 91(3):pp1059–1063.

Fujiwara, Y., and Konagaya, A. 1993. protein motif extraction using hidden markov model. In *Proceedings of Genome Informatics Workshop IV*, pp56-64.

Haussler, D.; Krogh, A.; Mian, I.; and Sjolander, K. 1993. protein modeling using hidden markov models: analysis of globins. *Proceedings of the Twenty-sixth Hawaii International Conference on System Science* I:pp792–802.

Konagaya, A., and Kondou, H. 1993. stochastic motif extraction using a genetic algorithm with the mdl principle. *Proceedings of the Twenty-sixth Hawaii International Conference on System Science* I:pp746–753.

Nakagawa, S. 1988. *Speech Recognition Using Stochastic Models.* Electronic Society of Information Communication.

Takami, J., and Sagayama, S. 1991. automatic generation of the hidden marcov network by successive state splitting. In *Proceedings of ICASSP.*