Hidden Markov Models and Iterative Aligners: Study of their Equivalence and Possibilities

Hidetoshi Tanaka and **Masato Ishikawa**
Institute for New Generation Computer Technology (ICOT)
1-4-28, Mita, Minato-ku, Tokyo 108 Japan
htanaka@icot.or.jp, ishikawa@icot.or.jp

Kiyoshi Asai
Electrotechnical Laboratory (ETL)
1-1-4, Umezono, Tsukuba
Ibaraki 305 Japan
asai@etl.go.jp

Akihiko Konagaya
C&C Systems Research Laboratories
NEC Corporation
4-1-1, Miyazaki, Miyamae-ku
Kawasaki, Kanagawa 216 Japan
konagaya@csl.cl.nec.co.jp

Abstract

There are many shared attributes between existing *iterative aligners* and *Hidden Markov Model* (HMM). A learning algorithm of HMM called *Viterbi* is the same as the iteration of DP-matching of iterative aligners. HMM aligners can use the result of an iterative aligner initially, incorporate the similarity score of amino acids, and apply the detailed gap cost systems to improve the matching accuracy. On the other hand, the iterative aligner can inherit the modeling capability of HMM, and provide the better representation of the proteins than *motifs*. In this paper, we present an overview of several iterative aligners which include the parallel iterative aligner of ICOT and the HMM aligner of Haussler's group. We compare the merits and shortcomings of these aligners. This comparison enables us to formulate a better, more advanced aligner through proper integration of the iterative technique and HMM technique.

Introduction

It is indispensable to align multiple protein sequences in order to understand the relationship among the protein function, the structure, and the amino acid sequence. The *functional motif* or the *structural motif* are found when the protein sequences of the same function or the same structure are multiply aligned. There are many methods to align them which include *iterative aligners* and *Hidden Markov Model* (HMM) aligners.

In order to improve the performance of the aligners and the quality of the result, several iterative aligners have been developed which include the parallel iterative aligner developed at ICOT (Ishikawa et al. 1992). This aligner is based on the Gotoh method (Gotoh 1992) with its parallel extension and *restricted partitioning* like the Barton-Sternberg method (Barton & Sternberg 1987). It quickly makes multiple sequence alignment of good score on parallel inference machines developed at ICOT.

The existing aligners however have two problems. One is about exploring an appropriate *gap costs* system. Both pairwise aligners and multiple aligners which use DP-matching algorithms (Needleman & Wunsch 1970) are suffered from what kind of score system to use. We have *Dayhoff PAM-250* (Dayhoff, Schwartz & Orcutt 1978) as a standard similarity score system but we have no standard way of estimating gap costs (Altschul 1989).

The other is about the representation of the result. From the resultant alignment of a certain protein set, we determine a specific consensus pattern called *motif* (Staden 1988) to represent the set. From the information retrieval point of view, its *recall* and *precision* (Frakes & Baeza-Yates 1992) is rather sufficient. For example, the motif database system, ProSite (Bairoch 1991), classifies each protein into five sorts on the relationship between each motif and proteins in Swiss-Prot: true-positive(true, for short), false-negative(missed), potential-hit, unknown, and false-positive(wrong), where false-negative proteins are true but missed by the motif matching while false-positive proteins are picked but wrong. A rough estimation is shown in Table 1. That estimates its recall by true/(true+missed) and its precision by true/(true+wrong). However, it seems insufficient for the biological practical use. There is a trade-off between

description complexity of patterns and improvement of the recall and precision.

Table 1: Estimation of Recall & Precision of Motifs

ProSite	true	missed	wrong	unk.	pot.
V.7	10787	260	571	75	630
V.9	12736	352	661	86	727

ProSite	recall	precision
V.7	97.6 %	94.9 %
V.9	97.3 %	95.0 %

In order to improve representation capability of the motif and to avoid description complexity, we propose to introduce probability distribution. There are several related works that employ probability in the motif representation, for example, probability with MDL criteria (Konagaya & Kondo 1993). Among them, we consider the Hidden Markov Model is the most suitable for representing the protein classification. It represents the protein set by the network whose nodes and arcs have probability distributions. Thus, it can be regarded as flexible multiple consensus patterns for the protein set.

The Hidden Markov Model can be easily applied to protein sequence alignment, as the DP-matching algorithm has been applied to it, by regarding sequences as speech. The DP-matching algorithm and HMM are popular tools among the researchers of speech recognition. In addition, the HMM employs several algorithms in its statistical modeling, including the *Viterbi* algorithm (Rabiner 1989) that is the same as the DP-matching. Thus, the HMM can be viewed as a concept that encompasses the DP-matching. Protein structure prediction from sequence (Asai & Hayamizu 1991)(Asai, Hayamizu & Onizuka 1993), and protein classification (White, Stulz & Smith 1991) are related works of biological information processing using HMM.

We take the *intelligent refiner* (Hirosawa 1993) into account to consider specific characteristics of single amino acid, such as cysteine(C) and histidine(H) in the zinc finger motif. The aligned sequences are refined by biological knowledge such as motifs in the ProSite. A new knowledge representation language *Quixote* using DOOD and CLP is proposed for representing such biological knowledge (Tanaka 1993). This paper however discusses the previous stage of the refinement with biological knowledge bases. We introduce no prior knowledge of the protein structure, function, or the characteristic of each amino acid at this stage.

This paper is organized as follows. Section overviews the variety of existing iterative aligners. Section shows an example of the HMM aligner. Section clarifies their relationship, the possibility of improving multiple sequence aligning algorithm, and appropriate representation of the alignment. And Section enumerates our future works.

Iterative Aligners

In this section, we show an overview on several iterative aligners. The purpose is to compare between these iterative aligners and the HMM aligner to be mentioned in Section .

Barton & Sternberg (1987)

Their algorithm to align N sequences is as follows:(Barton & Sternberg 1987)

(1) choose 2 out of the N sequences to do DP-matching between them, and obtain the result,
(2) choose 1 out of the rest $N-2$ sequences to do DP-matching with the last result,
(3) choose 1 out of the rest $N-3$ sequences to do DP-matching with the last result,
(4) ....
(5) do DP-matching between the last sequence and the last result, then
(6) obtain the final result.

The following refinements are done, if necessary.

(1) Choose the 1st sequence to do DP-matching with the rest of the latest result,
(2) choose the 2nd sequence to do DP-matching with the rest of the latest result,
(3) ...
(4) choose the N-th sequence to do DP-matching with the rest of the latest result.

When they do DP-matching between single sequence and plural sequences, they employ a specific score system for the plural sequences to do a kind of 2-way DP-matching, which we regard as the *profile DP-matching*. An example of the profile is shown in Figure 1.

This profile is a sequence position-specific scoring matrix composed of 21 columns (for 20 amino acids and gap cost) and N rows, where N is a length of *probe* (Gribskov, McLachlan & Eisenberg 1987). The probe is a bundle of aligned sequences. When we have a bundle of aligned sequences and a new sequence and want to align all the sequences, we can define the profile of the bundle and do DP-matching using scores of the profile instead of usual similarity scores, such as PAM-250. It is called the profile DP-matching.

Berger & Munson (1991)

Figure 2 shows a brief overview of the algorithm (Berger & Munson 1991).

(1) Let initial state be unaligned sequences,
(2) partition all sequences randomly into 2 groups to do profile DP-matching, and have the resultant alignment,
(3) repeat until the score of the alignment converges.

Pos	Probe	Consensus	Profile																							
		||	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	gap
1	E	G	V	L	V	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-2	9
2	L	L	S	P	L	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-1	9
3	V	V	V	V	V	2	2	-2	-2	2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-1	9
4	K	E	A	T	A	6	-2	5	6	-5	4	1	0	5	-2	0	3	3	3	1	3	6	0	-6	-4	9
5	A	P	L	P	P	6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5	-4	9
6	G	G	G	G	G	7	1	7	5	-6	15	-1	-3	0	-4	-3	4	3	2	-3	6	4	2	-11	-7	9
7	S	S	Q	E	D	4	-1	7	7	-6	7	2	-2	2	-3	-2	4	3	6	1	6	2	-1	-6	-5	9
8	S	S	T	P	S	4	4	2	2	-4	4	-1	0	2	-3	-2	2	7	0	1	10	6	0	-2	-4	9
9	V	L	V	A	V	5	0	-1	-1	3	1	-2	7	-2	7	6	-1	1	-1	-3	0	2	10	-5	-1	9
10	K	R	R	S	R	0	-1	1	1	-5	0	2	-2	8	-3	1	3	3	3	10	5	1	-2	7	-5	9
11	M	L	I	I	I	0	-2	-3	-2	7	-3	-3	11	-1	11	10	-2	-2	-1	-2	-2	1	9	-3	1	9
12	S	S	T	S	S	4	6	2	2	-3	5	-1	0	2	-3	-2	3	4	-1	1	12	6	0	0	-4	9
13	C	C	C	C	C	3	15	-5	-5	-1	2	-1	3	-5	-8	-6	-3	1	-6	-3	7	3	3	-13	10	9
14	K	S	Q	R	K	1	-2	3	3	-6	1	3	-2	7	-3	0	3	3	5	7	6	1	-2	2	-5	9
15	A	A	G	S	A	10	3	4	3	-5	8	-1	-1	1	-2	-1	3	4	1	-2	7	4	2	-6	-4	9
16	T	S	D	S	S	4	3	5	4	-5	6	0	0	2	-3	-2	4	3	1	1	9	6	0	-3	-4	9
17	G	G	S	Q	G	5	1	6	5	-6	9	4	-2	4	-3	-2	4	3	4	0	6	3	0	-6	-6	9
18	Y	F	L	S	F	-1	2	-4	-3	9	-3	0	4	-3	6	3	-1	-3	-3	-3	1	-1	2	7	7	9
19	T	T	R	L	T	1	-2	0	1	0	0	0	2	2	2	3	1	1	1	3	1	7	2	1	-2	9
20	F	F	.	L	F	-2	-3	-6	-4	10	-4	-1	6	-4	9	6	-3	-4	-4	-3	-2	-1	3	7	8	4
21	S	S	.	D	S	3	2	5	4	-4	5	0	-1	2	-3	-2	4	3	1	1	8	2	-1	-2	-3	4
22	S	.	.	S	S	2	3	1	1	-2	3	-1	0	1	-2	-1	2	2	0	1	8	2	0	1	-2	4
23	.	.	.	G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
24	.	.	.	D	D	1	-1	4	3	-2	2	1	0	1	-1	-1	2	1	2	0	1	1	0	-3	-1	4
25	.	.	.	G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
26	.	A	G	N	A	6	0	4	3	-4	6	1	-1	1	-2	-1	5	2	2	-1	3	3	1	-5	-3	4
27	Y	N	Y	T	Y	0	5	0	-1	5	-1	2	1	-1	0	-1	4	-3	-2	-2	0	3	0	3	6	4
28	E	D	D	Y	D	2	-2	9	8	-3	3	4	-1	1	-3	-2	5	-1	4	-1	1	1	-1	-6	0	9
29	L	M	A	L	L	3	-5	-3	-1	6	-1	-2	6	-1	10	10	-2	0	0	-2	-1	0	6	-1	0	9
30	Y	N	A	W	N	4	1	3	2	0	2	3	-1	1	-1	-1	8	0	1	-1	2	1	-1	-1	2	9
		:																								
48	S	G	N	S	S	4	3	5	3	-4	7	0	-2	2	-4	-3	6	3	1	0	10	3	0	-2	-4	9
49	S	S	N	Y	S	2	5	2	1	1	2	1	0	1	-2	-2	5	1	-1	0	8	1	-1	3	1	9

Figure 1: An Example of the Profile (Gribskov, McLachlan & Eisenberg 1987)

The sequences are randomly partitioned, so $2^{n-1}-1$ ways of partition may occur for n sequences. They ignore the column which contains any number of gaps using the Murata method (Murata, Richardson & Sussman 1985) in order to reduce computation. Iterative cycles of the Berger-Munson algorithm needs so much computation to take more than several hours to align practical scale of the alignment. Thus, parallel processing for this algorithm has been tried at ICOT (Ishikawa et al. 1992).

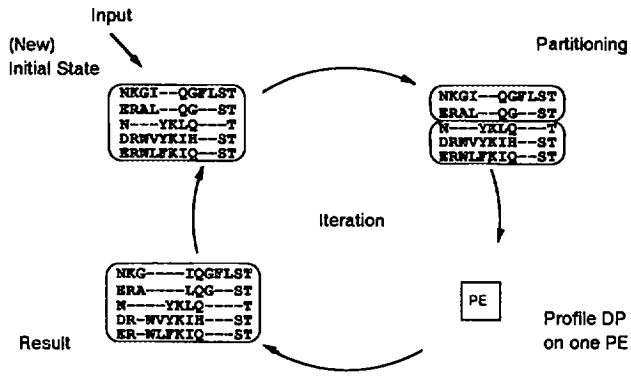

Figure 2: Berger-Munson Method

Gotoh (1992)

The similar algorithm as the Berger-Munson method, but Gotoh emphasizes the importance of the detailed gap costs (Gotoh 1992). Thus, this algorithm does not employ profile DP-matching so that it takes much time to execute every iteration cycle.

Before discussing gap costs, we should first define several terms. Let the gap cost of continuous n columns G_n be $O+(n-1)E$, we call O as *opening gap cost* and E as *extending gap cost.* Conventional simple algorithms use $O = E$ or $E = 0$.

The gap cost system of this method treats not only such differences between opening and extending gap costs. The cost of the i-th column G_i is $\sum^j \sum^k g_{ijk}$ where $g_{ijk} = 0$ if j-th and k-th sequence of the i-th column are both gaps or amino acids, otherwise $g_{ijk} = O$ or E.

ICOT (1992)

The similar algorithm as the Gotoh method, with its parallel extension (Ishikawa et al. 1992) and the restricted partitioning. Figure 3 shows a brief overview of the algorithm.

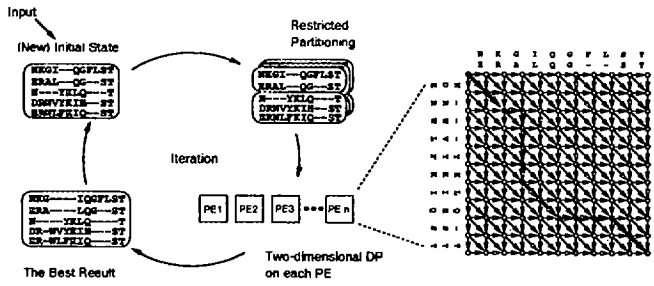

Figure 3: Parallel Extension & Restricted Partitioning of the Gotoh Method

Multiple branches of the search tree appeared at the restricted partitioning stage are evaluated in parallel with plural processing elements (PEs) in every iteration. Every possible partitioning into two groups of aligned sequences can be evaluated by 2-way DP-matching with the Gotoh gap cost system in parallel.

The restricted partitioning is an effective heuristic technique. N sequences are partitioned into k and $N-k$ according to the Berger-Munson method and the Gotoh method. Through the evaluation of the method, it is found that smaller k makes larger improvement in DP-matching score. Thus, we set $k=1$ or $k=2$ and call this heuristics as restricted partitioning. It resembles the Barton-Sternberg method in doing DP-matching between plural sequences and a few sequences.

Alignment using HMM

In this section, we show an HMM aligner to clarify the relationship between iterative aligners mentioned in

Section and the learning algorithm of the HMM aligner.

Haussler et al.(1993)

Their *Hidden Markov Network* (HMNet) is shown in the Figure 4, which is intuitively the simplest model for the multiple sequence alignment. The amount of protein sequence data is not enough for complicated HMNet to learn its probability distributions. Table 2 shows the amount of the sequences in representative protein databases. They are much fewer than the amount used in the speech recognition field. The complexity of the HMNet, which corresponds to the representation capability of the model, gets lower if we cannot prepare enough data.

Table 2: Statistics of Protein Databases

Databases	NRDB (1993)	PIR (1993)	Swiss-Prot V24 (1993)	PDB 1992
Proteins	61870	47234	28154	1146
Residues	17 M	13.8 M	9.5 M	-

NRDB: (Harris, States & Hunter 1993).

In Figure 4, columns of the HMNet roughly correspond to the columns in the alignment, though the number of the columns of the HMNet is generally fewer than the alignment. There are three kinds of nodes, d-node, i-node and m-node, which respectively correspond to deletion, insertion and matching of residues against the model. Only m-nodes have probability distributions, which correspond to the multiple consensus patterns with probability. This restriction is to reduce parameter of the model.

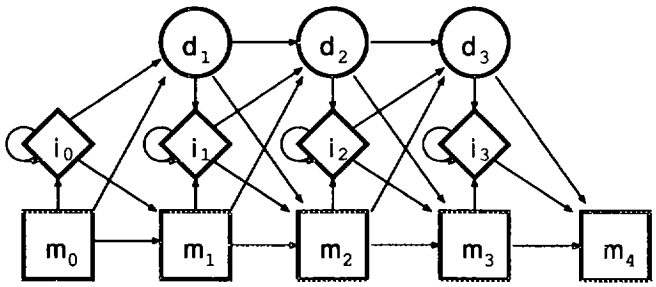

Figure 4: Hidden Markov Network of the Haussler (Haussler et al. 1993)

Their estimating algorithm is the Viterbi approximation (Haussler et al. 1993).

(1) Initially arcs from every node and every m-node have flat probability distributions,
(2) the most probable path through the model for each input protein sequence is determined by the Viterbi algorithm,
(3) probability is virtually a bit increased where the path is located,
(4) probability distribution of every node and arc is reestimated according to the sum of the virtual increase of probabilities when every sequence of the protein set are once input into the HMNet, and
(5) repeat determination of a path and reestimation of probability distributions, until it converges.

To achieve the optimal length of the model (the number of the m-nodes in Figure 4), the length is also reestimated by *model surgery*. If the frequency of located paths at a certain d-node is fewer than some threshold, or the frequency at a certain i-node is more than some threshold, the length of the model is reestimated and the model is retrained. To avoid the local optimization, linearly decreasing noise is added to the model during the beginning iterations, just as *simulated annealing* with the simplest annealing schedule. To avoid overfitting to the test data set, biases are introduced in the probability estimation process. They say this appears to be effective when large biases are given at the transition probability from i-nodes to m-nodes.

Evaluation

In this Section, we compare iterative aligners with the HMM aligner, examine their gap cost systems, and evaluate the HMM as representation of protein sets.

Comparison between Iterative Aligners and HMM

The HMM aligner resembles iterative aligners rather than n-way DP-matching algorithms (Carrillo & Lipman 1988), because the former have iterative improvement process like learning process in the HMM. Existing iterative aligners shown in Section and the learning algorithm of the Haussler HMM in Section employ the same algorithm for the same objects. The DP-matching in iterative aligners are called Viterbi approximation in the HMM. Both iterative and HMM aligners deal with columns of the alignment as objects. Hence, HMM is regarded as a meta-algorithm which describes how to apply DP-matching to the multiple sequence alignment, and how to model the aligned sequences. On the other hand, the learning algorithm of the Haussler HMM remains to be improved a lot, by introducing technique of iterative aligners.

Among iterative aligners, the Barton-Sternberg method have the highest similarity to the Haussler HMM. It has a one-by-one DP-matching process and refinement processes, which correspond respectively to the path determining process and the reestimation process of the Haussler HMM, though the path determining by Viterbi approximation is not one-by-one process practically. It employs the profile DP-matching, whose treatment of gap costs is similar to the HMM, as both are independent from horizontal positions of the residues.

Comparison of Gap Cost Systems

Most importantly, HMM treats gap costs much easier than DP-matching algorithms. We have struggled to introduce detailed gap cost system into various DP-matching based multiple alignment system, but found no proper way to do so. HMM can also treats similarity score system such as PAM-250. We can introduce it into the probability distributions of m-nodes either previously or after the HMM is estimated.

The Haussler HMNet is shown again in Figure 5. In this HMM, opening gaps and extending gaps correspond to the probability of the arc. Arrows between m-nodes are normal transitions, and from d-nodes and i-nodes to m-nodes correspond to the end of gaps. Obviously, insertion and deletion gap costs as well as gap costs at different nodes are independently determined in the Haussler HMM while they should be the same cost in the conventional DP-matching algorithms. There was no elegant way to treat such locality in the conventional DP-matching. The HMM naturally enables us to distinguish important conservative regions which contain consensus patterns from other regions of sequences in the alignment process.

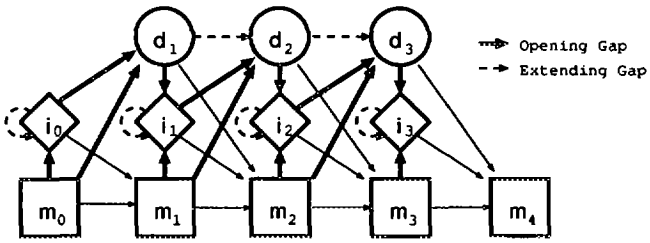

Figure 5: Gap costs in the Haussler HMM

HMM realizes natural implementation of the gap cost $G_N = P_O + (N-1)P_E$, which is a standard in DP-matching. (Let the transition probability corresponding to an opening gap be A_O and an extending gap be A_E, and their costs be $P_O = -\log A_O$ and $P_E = -\log A_E$, then, the gap cost of length N in usual HMM is $G_N = -\log A_O A_E{}^{(N-1)} = P_O + (N-1)P_E$.)

The feature specific to the Gotoh method is its detailed gap costs, which is also employed by the ICOT method. It requires some integrity of gap costs between insertion and deletion or among columns of the HMNet. We consider it can be realized by *duration modeling* (Rabiner 1989). It is an implementation technique to substitute the natural gap cost $G(N)$ by any $f(N)$.

Isolated residue problem also seems to be solved by introducing the probability. It is not essential but annoying, that a certain amino acid matches to a specific column to separate a long gap into two, instead that it should be put at the either end of the gap.

HMM as Representation of Protein Sets

The functional and structural motifs whose functions or structures are chemically or physically cleared would survive. We however consider the motifs as representation of the protein set should be replaced by the HMM. We have a new sequence and want to classify it among the existing protein sets. But we have no HMM representation of the set, then, we should re-align the sequence with every protein set, or extract motifs previously and search them against the sequence. The former costs expensive, no matter we have only the resultant alignment or we have its profile score, because it needs DP-matching which costs $O(N^2)$. The latter has a trade-off between the complexity of the motifs and the recall and precision rate. The HMM representation contains the probability distribution that relaxes the complexity and gains high rates.

However, there is another trade-off between the representation capability and the calculation speed. There are no proper indexing systems for motif searching, but motif scanning is almost the same algorithm as the regular expression search, which has long been researched so that many rapid and excellent methods have been well-developed. Whereas HMM has relatively less extensive work on retrieval in speech recognition, because retrieving the optimal HMM to know its classification is not the goal of that field.

Invisibility is another problem. Motifs are described by regular-expression like description, which we can read and understand easily. The HMM is in that sense rather invisible for us.

Future Work

In this section, we enumerate our future work according to the evaluation of the HMM as an aligner and a representation of a protein set shown in Section .

HMM as a Multiple Aligner

We introduce HMM concepts into the multiple aligner in three ways:

(1) mapping detailed gap cost system of iterative aligners to HMM,
(2) introducing similarity score into its probability distribution, and
(3) using sequences aligned by iterative aligners as the initial state.

Introduction of the similarity score could be instantly experimented. It is expected to have a good result, because it contributes to the reduction of parameters and the introduction of similarity, and because it has no defects from the biological point of view. We can start experiments from the conventional standard score PAM-250. Aligned initial state is also expected a good result. Mapping gap costs could be experimented soon, though there remain issues to be settled (see Section).

HMM as a Representation

We introduce HMM concepts into the representation of the protein set by providing an HMM viewer to exhibit HMM as well as motifs, and an HMM inverted file to make search faster. For biologists' practical use of the protein set representation, we should provide some expression of HMM in text and/or in display. This expression is also used as comparison with the structural or functional motifs found by chemical and physical means. As for inverted file or index for HMM, we may build HMMs hierarchically along with, for example, the protein hierarchy such as cytochromes, cytochrome c, and cytochrome c555.

Concluding Remarks

The learning algorithm of HMM is the same as the iteration of DP-matching. HMM aligners can use the result of an iterative aligner initially, incorporate the similarity scores such as PAM-250, and apply the detailed gap cost systems to improve the matching accuracy. On the other hand, the iterative aligner can inherit the modeling and representational capability of HMM. We present an overview of several iterative aligners and an HMM aligner to compare the merits and shortcomings of these aligners. This comparison enables us to formulate a better, more advanced aligner through proper integration of the iterative technique and HMM technique.

The HMM is so suitable for representing the protein classification. It is a natural representation of multiple consensus patterns with probability distribution. Motif is useful for representing chemical and physical analysis, and for browsing by biologists. Motifs and HMM will cooperate to represent protein classifications.

Acknowledgments

The authors wish to thank Kentaro Onizuka and Stephen T.C. Wong for their valuable comments on earlier versions of this paper. The authors also thank Katsumi Nitta, Masaki Hoshida, Makoto Hirosawa, Tomoyuki Toya and the people in the GIP project at ICOT for fruitful discussions and great efforts to realize systems.

References

Altschul, S.F. 1989. Gap Costs for Multiple Sequence Alignment. *J. Theor. Biol.* 138:297-309.

Asai,K.; Hayamizu,S.; and Handa,K. 1991. Secondary Structure Prediction by Hidden Markov Model. *CABIOS* Forthcoming.

Asai,K.; Hayamizu,S.; and Onizuka,K. 1993. HMM with Protein Structure Grammar. In Proceedings of the Twenty-sixth Annual Hawaii International Conference on System Sciences, 783-791. Los Alamitos, Calif.: IEEE Computer Society Press.

Bairoch, A. 1991. PROSITE: A Dictionary of Protein Sites and Patterns, User's Manual.

Barton, G.J.; and Sternberg, M.J. 1987. A Strategy for the Rapid Multiple Alignment of Protein Sequences: Confidence Levels from Tertiary Structure Comparisons. *J. Mol. Biol.* 198:327-337.

Berger, M.P.; and Munson, P.J. 1991. A Novel Randomized Iterative Strategy for Aligning Multiple Protein Sequences. *CABIOS* 7(4):479-484.

Carrillo, H.; and Lipman, D. 1988. The Multiple Sequence Alignment Problem in Biology. *J. Appl. Math.* 48:1073-1082.

Dayhoff, M.O.; Schwartz, R.M.; and Orcutt, B.C. 1978. A Model of Evolutionary Change in Protein. *Atlas of Protein Sequence and Structure* 5(3):345-352. Washington D.C.: Nat. Biomed. Res. Found.

Frakes, W.B., and Baeza-Yates, R. eds. 1992. *Information Retrieval, Data Structures & Algorithms.* Prentice Hall.

Gotoh, O. 1992. Optimal Alignment between Groups of Sequences and its Application to Multiple Sequence Alignment. *CABIOS* Forthcoming.

Gribskov, M.; McLachlan, A.D.; and Eisenberg, D. 1987. Profile Analysis: Detection of Distantly Related Proteins. *Proc. Natl. Acad. Sci. USA* 84:4355-4358.

Harris, N.L.; States, D.J.; and Hunter, L. 1993. ClassX: A Browsing Tool for Protein Sequence Megaclassification. In Proceedings of the Twenty-sixth Annual Hawaii International Conference on System Sciences, 554-563. Los Alamitos, Calif.: IEEE Computer Society Press.

Haussler, D.; Krogh, A.; Mian, I.S.; and Sjölander, K. 1993. Protein Modeling using Hidden Markov Models: Analysis of Globins. In Proceedings of the Twenty-sixth Annual Hawaii International Conference on System Sciences, 792-802. Los Alamitos, Calif.: IEEE Computer Society Press.

Hirosawa, M.; Hoshida, M.; and Ishikawa, M. 1993. Protein Multiple Sequence Alignment using Knowledge. In Proceedings of the Twenty-sixth Annual Hawaii International Conference on System Sciences, 803-812. Los Alamitos, Calif.: IEEE Computer Society Press.

Ishikawa, I.; Hoshida, M.; Hirosawa, M.; Toya, T.; Onizuka, K.; and Nitta, K. 1992. Protein Sequence Analysis by Parallel Inference Machine. In Proceedings of the International Conference on Fifth Generation Computer Systems 1992, 294-299. Tokyo, Japan: Ohmsha, Ltd.

Konagaya, A.; and Kondo, H. 1993. Stochastic Motif Extraction using a Genetic Algorithm with the MDL principle. In Proceedings of the Twenty-sixth Annual Hawaii International Conference on System Sciences, 746-753. Los Alamitos, Calif.: IEEE Computer Society Press.

Murata, M.; Richardson, J.S.; and Sussman, J.L. 1985. Simultaneous Comparison of Three Protein Sequences. *Proc. Natl. Acad. Sci. USA* 82:7657.

Needleman, S.B.; and Wunsch, C.D. 1970. A General Method Applicable to the Search for Similarities in the Amino Acid Sequences of Two Proteins. *J. Mol. Biol.* 48:443-453.

Rabiner, L.R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE* 77(2):257-286.

Staden, R. 1988. Methods to Define and Locate Patterns of Motifs in Sequences. *CABIOS* 4(1):53-60.

Tanaka, H. 1993. A Private Knowledge Base for Molecular Biological Research. In Proceedings of the Twenty-sixth Annual Hawaii International Conference on System Sciences, 844-852. Los Alamitos, Calif.: IEEE Computer Society Press.

White, J.V.; Stulz, C.M.; and Smith, T.F. 1991. Protein Classification by Nonlinear Optimal Filtering of Amino Acid Sequences *IEEE Transactions on Signal Processing* Forthcoming.