Profiling Protein Families from Partially Aligned Sequences*

Saikat Mukherjee
Siemens Corporate Research
saikat.mukherjee@siemens.com

Chang Zhao I.V. Ramakrishnan
Stony Brook University
{changz,ram}@cs.sunysb.edu

Abstract

Profile Hidden Markov Models (PHMMs) are recognized as powerful computational vehicles for homology search of protein sequences. Extant PHMM training approaches either use completely unaligned or aligned sequences. The PHMMs resulting from these two training approaches present contrasting tradeoffs w.r.t. alignment information and the accuracy of the search outcome. This paper describes a PHMM based technique for modeling protein families from partially aligned sequences. By exploiting the observation that partially aligned sequences give rise to independent subsequences, PHMMs corresponding to these subsequences are composed to build PHMMs for the entire sequences. An interesting aspect of the technique is that it gives rise to a family of PHMMs which are parameterized w.r.t. the alignment information. We present experimental comparison of the performance of our technique against several state of the art homology detection methods.

1 Introduction

The success of genomic work on various species has resulted in an enormous multitude of biological sequence information. This has created a rich research area centered around the development of automated techniques for analysis of these sequences. An effective means of understanding the characteristics of a new biological polymer from its sequence is through homology – whereby the sequence is compared to other similar sequences with known rich biological information. Profile hidden Markov Model[2, 5] has been proven to be able to detect remote homology.

The two dominant approaches for training PHMMs differ mainly in the way training sequences are utilized. At one extreme is training from *completely aligned* sequences where all the residues in every sequence are mapped to a column representation taking into account insertions and deletions. In contrast, (inexpensive) training from *completely unaligned* sequences uses no such information. Not surprisingly, PHMMs trained from completely aligned sequences (which we will refer to as A_PHMMs) have been shown to identify remote homologs with a much higher degree of accuracy than those trained from unaligned sequences (which we will refer to as U_PHMMs). However, producing the information about alignments is a labor intensive process involving expensive structural analysis of entire sequences. The contrasting trade-offs at the two ends of the alignment spectrum gives rise to the question: Can we develop techniques for learning profile PHMMs that trade the accuracy of remote homolog identification for alignment information? Using the notion of *partially aligned* sequences where only parts of sequences are aligned against each other, we formulate this problem as one of estimating PHMM parameters from such sequences. We will refer to PHMMs trained with such partially labeled sequences as P_PHMMs.

The essence of our approach for training PHMMs from partially aligned sequences (referred to as P_PHMM from now on) rests on the observation that a consecutive string of unaligned residues between two aligned residues can be generated only from the sequence of states lying between the match states for the aligned residues in the P_PHMM structure. Based on this observation, the algorithm decomposes P_PHMM into submodels whose parameters are separately estimated and then composed together to produce the original P_PHMM parameters. The technique is *parameterized* w.r.t. the alignment information in the sense that by varying the alignment information we can estimate the parameters of PHMMs spanning the entire spectrum from aligned PHMM at one end to unaligned PHMM (U_PHMM) at the other end.

The idea of combining PHMMS has been explored by MetaMEME[4]. However, our approach uniformly models both motif and non-motif regions as full PHMMs with match, insert, and delete states. This leads to more precise results especially when motifs do not cover significant portions of the sequences. Another closely related work is TCoffee[7] which can be used to generate a multiple alignment (from which a family model can be learned) from partial alignments.

The rest of the paper is organized as follows: In Section 2, we present algorithmic details of our technique for building P_PHMMs. Section 3 presents experimental results on the performance of P_PHMM. Section 4 concludes the paper.

*We thank Dr. Subramanyam Swaminathan for insightful discussions on this work. This research is supported by U.S. Army Medical Research Acquisition Activity Contract DAMD17-03-1-0520 and New York State Department of Health Contract C020593.

Figure 1: Partial alignment information for ten ig sequences

2 Partial Alignment Profiling

Building P_PHMMs rests on the use of *partial alignment* information to *decompose* a PHMM structure into submodels and *compose* parameters computed from these submodels into the PHMM's parameters.

Partially Aligned Sequences: In a set of partially aligned sequences, alignment information is known only for a subsequence of residues in every individual sequence in the set. C_1, C_2, and C_3 in Figure 1 show three aligned columns in the ten sequences of the ig family. The alignment C_1 spans the residues $A, V, L, I, L, A, K, V, M$, and A in the ten sequences respectively and is illustrated by the leftmost solid line. Similarly, the alignment C_3 spans the Y residues in each of the ten sequences as indicated by the rightmost solid line. As illustrated in C_2, where the residues S, D, F, T, and D in only the last five sequences are aligned, it is not necessary that an alignment information has to cover all the sequences in the set. In the event of alignment being known for all the residues in every sequence, partial alignment collapses to complete alignment while total absence of any alignment information reduces to a set of unaligned sequences.

We have used the simple heuristic of taking the average length of the sequences to estimate the model length. For instance, for the ten ig family members in Figure 1, the model length computed by averaging over the size of the ten sequences is 74. By the definition of alignment, all residues aligned at a particular column are generated from the same state in the PHMM. We estimate this state by averaging over the positions of the residues, belonging to the alignment, in their corresponding sequences. For instance, for the alignment C_1 in Figure 1, the mean position where a residue in the alignment occurs in a sequence is 12. Consequently, all the ten residues in C_1 are generated from the match state M_{12}. Similarly, the ten residues in C_3 and the five residues in C_2 are generated from the match states M_{65} and M_{21} respectively.

Model Decomposition: The key to using partial alignment information for estimating PHMM parameters is the observation that a substring of unaligned residues between any two aligned residues can only be generated from the sequence of states in model positions between those corresponding to the aligned residues. For instance, in the first sequence 1LTK in Figure 1, the residues A (C_1) and Y (C_3) belong to match states M_{12} and M_{65}. The substring of unaligned symbols from R to K between the two aligned residues can be generated only from states in model positions 13 to 64 and the insert state I_{12}. This observation lets us decompose the PHMM structure into submodels where each submodel generates substrings from the original sequence. In what follows, we have *ignored gaps* in alignment information for simplicity of exposition of our technique.

In our decomposition framework, aligned residues are generated from singleton match states while substrings of unaligned residues are generated from PHMMs consisting of states in sequences of consecutive positions in the original model. We construct these PHMMs, or submodels, from the appropriate states in the original model and add begin and end states to complete the submodel structure. The PHMM P_1 in Figure 2 illustrates an example submodel. During decomposition, for a sequence with aligned residues α_n, α_m generated at match states M_i, M_j respectively and with the intermediate unaligned substring $\alpha_{n+1} \cdots \alpha_{m-1}$ generated from the submodel P, transitions are created from M_i to P and from P to M_j. In the event of consecutive aligned residues (i.e. $\alpha_m = \alpha_{n+1}$), the submodel P

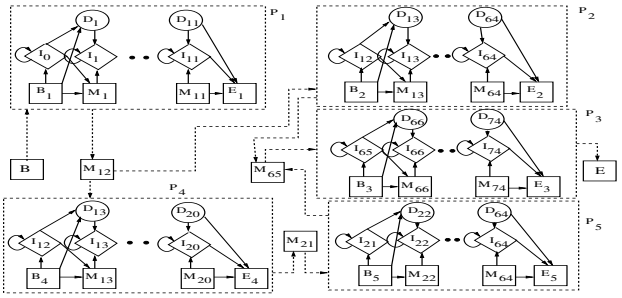


Figure 2: Decomposition of a 74 length PHMM structure using the partially aligned sequences in Figure 1

does not exist and M_i directly transitions to M_j. Figure 2 illustrates the decomposition on a PHMM structure with model length 74 using the ten partially aligned sequences of the ig family in Figure 1.

Parameter Composition: The essence of our composition technique is to estimate the original PHMM parameters from expectations of transition and emission events computed from submodels and singleton match states.

Recall that a submodel generates a set of unaligned residue substrings. For instance, the submodel P_1 in Figure 2 generates the first eleven residues of all the ten ig family sequences shown in Figure 1. This allows individual submodel parameters to be estimated by Baum-Welch training.

The singleton match states corresponding to aligned columns generate a set of residues. For instance, in Figure 2, M_{65} emits only the Y residue while M_{12} emits the residues A, V, L, I, K, and M. The emission probabilities of residues in these match states are estimated by smoothed maximum likelihood frequency counting. Also, transition probabilities between submodels and neighboring match states and vice-versa are estimated using a smoothed maximum likelihood approach. For instance, in Figure 2, if n_{M_{12},P_2} and n_{M_{12},P_4} denote the number of sequences where transitions from M_{12} to P_2 and from M_{12} to P_4 occur respectively, the probability of transition from M_{12} to P_2, $p_{M_{12},P}$, is computed as $\frac{n_{M_{12},P_2}+1}{n_{M_{12},P_2}+n_{M_{12},P_4}+2}$.

The partial alignment information in a sequence can be such that:

1. Alignment occurs at the match states M_k and M_l, where $k < i$ and $j < l$, for the residues α_n and α_m respectively.
2. Alignment occurs at the match state M_i for the residue α_n but not at M_{i+1}.
3. Alignment occurs at the match state M_{i+1} for the residue α_m but not at M_i (the converse of the above).
4. Alignment occurs at both M_i and M_{i+1} for residues α_n and α_{n+1} respectively.

Given a sequence, these four scenarios influence the computation of transition expectations for the three kinds of states in a PHMM. Let A_{S_1,S_2} denote the transition expectation between state S_1 and S_2. Apparently scenario 4 only contributes to $A_{M_i,M_{i+1}}$ which in this case is just the count of the number of times a transition is made between the singleton match states M_i and M_{i+1}. For the other three scenarios, Table 1 summarizes how the transition expectations are estimated.

Scenario	1	2	3
$A_{D_i,D_{i+1}}$	BW	N/A	N/A
A_{D_i,I_i}	BW	N/A	BW
$A_{D_i,M_{i+1}}$	BW	N/A	$A^P_{D_i,E_P} \times p_{P,M_{i+1}}$
A_{I_i,I_i}	BW	N/A	N/A
$A_{I_i,D_{i+1}}$	BW	N/A	N/A
$A_{I_i,M_{i+1}}$	BW	N/A	$A^P_{I_i,E_P} \times p_{P,M_{i+1}}$
A_{M_i,I_i}	BW	$p_{M_i,P} \times A^P_{B_P,I_i}$	BW
$A_{M_i,D_{i+1}}$	BW	$p_{M_i,P} \times A^P_{B_P,D_{i+1}}$	N/A
$A_{M_i,M_{i+1}}$	BW	$p_{M_i,P} \times A^P_{B_P,M_{i+1}}$	$A^P_{M_i,E_P} \times p_{P,M_{i+1}}$

Table 1: Transition Expectations

In Table 1, all entries marked by 'BW' means that the expectation is estimated from Baum-Welch on the appropriate submodel. For example, the expectation $A_{D_i,D_{i+1}}$ from delete state D_i to D_{i+1} for scenario 1 is given by $\sum_{t=n+1}^{t=m-1} \xi_t(D_i, D_{i+1})$.

Scenario 2 and 3 require considering a neighboring singleton match state. Let us work out scenario 3 for $A_{D_i,M_{i+1}}$. In such a situation, D_i makes a transition to the end state E_P of the submodel P which generates the unaligned substring preceding the aligned residue in M_{i+1}. Thus $A_{D_i,M_{i+1}}$ is estimated as $A^P_{D_i,E_P} \times p_{P,M_{i+1}}$, where $p_{P,M_{i+1}}$ is the probability of transition between P and M_{i+1}.

Finally, the sum of the expectations for any event over all the sequences are used to estimate its probability using a smoothed maximum likelihood technique. For instance, the probability of transition between M_i, M_{i+1} is given by:

$p_{M_i,M_{i+1}} = \frac{\sum A_{M_i,M_{i+1}}+1}{\sum A_{M_i,M_{i+1}}+\sum A_{M_i,I_i}+\sum A_{M_i,D_{i+1}}+3}$,

where the summation denotes the cumulative value of the expectation over all the sequences.

Emission expectations of residues in states are estimated from submodels, by Baum-Welch, and from singleton match states by frequency counting. Smoothed maximum likelihood is used to compute the emission probabilities from these expectations.

3 Experimental Results

Experiments were conducted to compare the performance of P_PHMM against vanilla PHMM (U_PHMM), SAM which is a state of the art PHMM tool, an advanced multiple alignment tool TCoffee, and metaMEME. The effect of varying training set size as

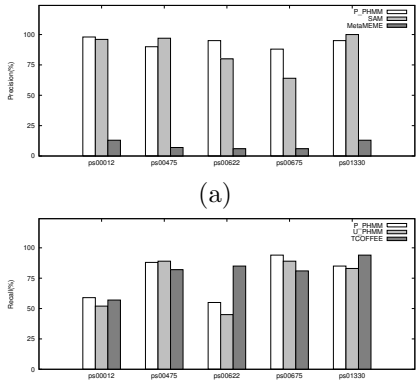
ID	M	P_PHMM		U_PHMM		SAM		TCOF.		MME.		RE	
		T	F	T	F	T	F	T	F	T	F	T	F
ps00012	221	130	3	115	0	165	6	127	2	139	967	169	143
ps00475	80	70	8	71	6	72	2	66	0	60	851	57	11
ps00622	100	55	3	45	0	85	21	85	0	84	1225	77	4
ps00675	91	86	11	81	19	86	49	73	1	86	1412	70	138
ps01330	96	82	5	80	2	90	0	90	0	24	164	59	0

(a)

ID	Members	Aligned Cols
a.1.1.2	60	186
b.1.1.2	59	122
b.34.2.1	26	176
c.47.1.5	31	101
d.169.1.1	28	173

(b)

Figure 3: (a)Experimental data with 15% training set on the 5 Prosite families (b)The 5 SCOP families



(a)

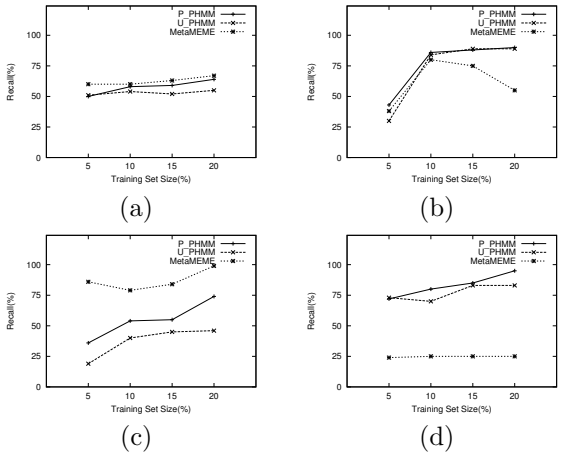
(b)

Figure 4: (a) P_PHMM precision against SAM and metaMEME (b) P_PHMM recall against U_PHMM and TCoffee

well as alignment information on the performance were also investigated.

Datasets: Regular expression based signature information available for families in the PROSITE database[3] were used to generate partially aligned sequences. Matches of a family's signature in sequences which belong to it constitute the partial alignment information for these sequences. To demonstrate the effectiveness of our technique in homology identification, 5 PROSITE families, each having at least 50 members, were chosen where RE-based pattern signatures were not very effective in identifying family members. The first column in Figure 3(a) shows the families used while the second column shows the number of members of each family in the Swiss-Prot database [1]. The models trained with P_PHMM, U_PHMM, SAM, and TCoffee were used with hmmsearch of HMMER [2] to detect homologs in Swiss-Prot while for metaMEME its own search tool, mhmms, was used. Default cutoff values were used in both the cases.

Recall and Precision: Figure 3(a) tabulates the results of the experiments for the five models on the five families using 15% of the members of each family as the training set. The columns T and F for each model reflect the number of true and false positives respectively in the test set. Figure 4 summarizes the results w.r.t. recall and precision. Observe from Figure 4(a) that the precision of P_PHMM is significantly better than metaMEME for all the families. The recall of P_PHMM is better than metaMEME for 3 of the 5 families as shown in Figure 3(a). The precision of P_PHMM is significantly better than SAM for ps00622 and ps00675 while being comparable for the other 3 families. Figure 4(b) illustrates the recall of P_PHMM against U_PHMM and TCoffee. P_PHMM has better recall for 4 of the families compared to U_PHMM and, apart from ps00622, has better or similar recall compared to TCoffee.



(a) (b) (c) (d)

Figure 5: Comparing recall of P_PHMM with U_PHMM and metaMEME for (a) ps00012, (b) ps00475, (c) ps00622, and (d) ps01330

Effect of varying training set: A desirable property of any supervised learning algorithm is the improvement in performance with increased training. Figures 5 shows the change in recall with increasing training set size for P_PHMM, U_PHMM, and metaMEME. Observe that for all the four families the recall of P_PHMM increases with training set size. In contrast, vanilla PHMM or U_PHMM does not always show a increase as evident in ps00012 and ps01330. This is even more true for metaMEME which, in spite of having similar recall numbers as P_PHMM in Figure 3(a), does not demonstrate better performance with more training.

Effect of varying alignment information (SCOP): The parameterized nature of the P_PHMM algorithm was borne out by experiments conducted on families from the SCOP [6] database. The SCOP database

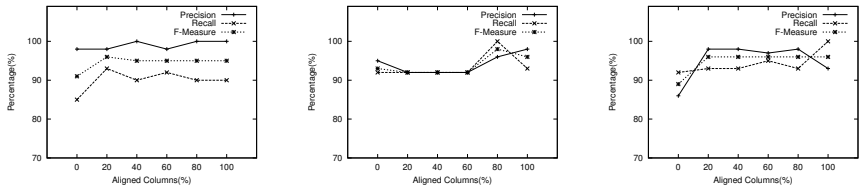


Figure 6: Impact on P_PHMM performance of varying alignment information for the SCOP families (a) a1.1.2, (b) b1.1.2, and (c) b.34.2.1

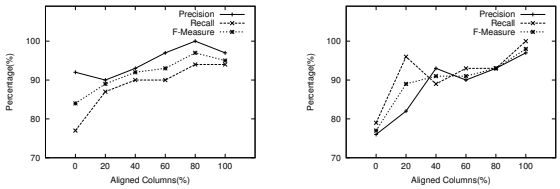


Figure 7: Impact on P_PHMM performance of varying alignment information for the SCOP families (a) c.47.1.5, and (b) d.169.1.1

provides detailed and comprehensive description of the structural and evolutionary relationship between proteins. For the purposes of our experiments, we selected 5 SCOP families each having at least 25 members. Multiple alignment of these families was derived from the PALI [8] database which provides alignments of proteins in the SCOP database. Column 1 in Figure 3(b) lists the ids of these 5 families, while Columns 2 and 3 show the number of family members and the number of aligned columns in their multiple alignment.

P_PHMMs were trained for each of these 5 families. The training set size, for each family, was fixed at a randomly chosen set of 25% of its total members. The amount of alignment information was successively varied from the use of 0% (completely unaligned), to 20%, 40%, 60%, and 80% of the number of aligned columns in the multiple alignment of the family. The test set for each of these families consisted of all the 5179 domains from all the 1029 families in PALI release 2.3. Recall, precision, and F-measure of homology detection were calculated for them. Figure 6 and Figure 7 graphically illustrates the impact on the three metrics with varying alignment information on all the 5 families. While all the 5 families show increase in the values of the three metrics with alignment information, this is especially perceptible in the SCOP families *c*.47.1.5 and *d*.169.1.1 in Figure 7(a) and (b) respectively.

4 Discussions

In this paper, we proposed a parameterized technique for learning PHMMs from partially aligned sequences. Our technique was based upon decomposing a PHMM structure into submodels and composing these submodels' parameters into that of the PHMM.

Usually, PHMM parameters are learned with the Baum-Welch algorithm from unaligned sequences. Note that it is non-trivial to modify Baum-Welch to handle partial alignment information. Baum-Welch is defined in terms of a pair of algorithms which are formulated in a greedy, recursive manner without any lookahead capability. Consequently, incorporating alignment information at a current position for residues which occur after it in the sequence is difficult. Considering such information is necessary to restrict the assignment of expectation values to valid states only. Incorporating other sources of partial alignment information, such as PSSMs, into our framework is a topic worth exploring.

References

[1] A. Bairoch and B. Boeckmann. The swiss-prot protein sequence data bank. *Nucleic Acids Res.*, 20:2019–2022, 1992.

[2] S. R. Eddy. Profile hidden markov models. *Bioinformatics*, 14:755–763, 1998.

[3] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. Sigrist, K. Hofmann, and A. Bairoch. The prosite database, its status in 2002. *Nucleic Acids Res.*, 30:235–238, 2002.

[4] W. Grundy, T. Bailey, C. Elkan, and M. Baker. Meta-meme: Motif-based hidden markov models of protein families. *Computer Applications in Biosciences*, 13(4):397–406, 1997.

[5] K. Karplus, C. Barrett, and R. Hugher. Hidden markov models for detecting remote protein homologies. *Bioinformatics*, 14:846–856, 1999.

[6] A. Murzin, S. Brenner, T. Hubbard, and C. Chothia. Scop: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247:536–540, 1995.

[7] C. Notredame, D. Higgins, and J. Heringa. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, 302:205–217, 2000.

[8] S.Balaji, S. Sujatha, S. Kumar, and N. Srinivasan. Pali-a database of alignments and phylogeny of homologous protein structures. *Nucleic Acids Res.*, 29:61–65, 2001.