Pacific Symposium on Biocomputing 4:162-170 (1999)

THE EFFECTS OF ORDERED-SERIES-OF-MOTIFS ANCHORING AND SUB-CLASS MODELING ON THE GENERATION OF HMMs REPRESENTING HIGHLY DIVERGENT PROTEIN SEQUENCES

M. A. MCCLURE, J. KOWALSKI
Department of Biological Sciences
UNLV, Las Vegas, NV 89129

Hidden Markov Models (HMMs) provide a flexible method for representing protein sequence data. Highly divergent data require a more complex approach to HMM generation than previously demonstrated. We describe a strategy of motif anchoring and sub-class modeling that aids in the construction of more informative HMMs as determined by a new algorithm called a stability measure.

1 Introduction

The genomes of RNA-based life-forms (e.g., HIV, Ebola and Measles) exist as quasi-species, with accompanying mutant clouds, due to the rapid rate at which RNA genomes can replicate and accumulate errors. [Domingo and Holland, 1997] These mutated RNA genomes provide us with a highly divergent set of co-linear genes encoding a variety of enzymatic and structural proteins. Many of the relationships among these protein sequences fail statistical criteria for homology, although all biological and biochemical data support common ancestry. We define such proteins as functionally equivalent "relatives " in contrast to those members of the set which are clearly homologous (usually greater than 25% identical). When proteins are this highly divergent, the regions of common residues, the ordered-series-of-motifs (OSM), are those that contribute to the function or structural integrity of the protein. [McClure, 1991]

The correct identification of these strings of common sub-sequences or OSM among a set of protein sequences is the first step in multiple sequence alignment. [McClure, et al., 1994] The second step requires the alignment of regions between the functionally selected OSM. The motif-intervening-regions (MIRs) are less constrained by the functional selection operating on the OSM. The MIRs, however, can be constrained by selection pressures specific to sub-classes of the sequence set and often change more rapidly relative to the OSM. MIRs can vary widely in size, and amino acid composition.

To access the maximum information contained in primary structure data both the OSM and MIRs must be aligned as precisely as possible. The OSM defines a pattern among the sequences that allows the possibility of common function and ancestry. These patterns populate motif databases. The MIRs can define sub-class

functional specificities and additional sub-class motifs. These regions contain information important to the reconstruction of the phylogenetic history of the protein sequences. All positions in the alignment provide data that can be used to test a wide variety of evolutionary hypotheses regarding gene and genome construction. Automated generation of a multiple alignment of large numbers of highly divergent homologous and functionally equivalent protein sequences remains a challenge in the field of bioinformatics.

In the studies initiated here we explore a method of incorporating the OSM information, *a priori*, using the Hidden Markov (HMM) approach [Rabiner, 1989] to model highly divergent protein sequence data. [Baldi, et al., 1994, Fujiwara, et al., 1994, Krogh, et al., 1994, Eddy, 1995, Hughey and Krogh, 1996] An HMM is essentially a stochastic production model consisting of a linear series of nodes. Each node contains the observation probabilities for match and insert states, and the transition probabilities between match, insert and delete states. The SAM 2.0 HMM method, used in this study, implements the full Baum-Welch expectation maximization algorithm with the injection of noise to avoid local optima. The Baum-Welch algorithm guarantees the likelihood of the model will increase with each training iteration. [Krogh, et al., 1994] Sequences are then aligned to the model using the SAM implementation of the Viterbi algorithm. [Rabiner, 1989] The advantages of the HMM approach are: 1) knowledge of the phylogenetic history or pairwise ordering is not required, 2) indel penalties are variable and position dependent, 3) the model can provide information regarding stochastic and selected features of a protein family, 4) information can be incorporated into the model *a priori*, and 5) the computation cost of aligning a set of sequences to an HMM is linearly proportional to the number of sequences to be aligned.

In earlier studies we explored some of the parameters involved in building HMMs for distantly related protein sequences. [McClure and Raman, 1995, McClure, et al., 1996] It was demonstrated that HMM approaches perform as well as or better than traditional dynamic programming algorithms in identifying the OSM in four benchmark protein families. Not even the HMM approaches, however, can correctly identify the complete OSM in the most distantly related members in two of the protein families. [compare data from McClure, et al., 1994 with McClure, et al., 1996] The correct identification of the OSM that defines membership in a specific protein family or class is the first criterion for constructing a meaningful HMM representing the sequence data. [see paper by Hudak and McClure submitted to this proceedings]

We are interested in the construction of HMMs that adequately reflect the evolutionary relationship for the entire length of all sequences of a given protein class. In our attempts to construct a HMM representing over 500 unique reverse transcriptase (RT) sequences found in the retroid family, we developed a strategy of HMM construction based on OSM-anchoring and sub-class modeling. These studies

generated multiple alignments from numerous HMMs requiring an automated scoring method to assess the ability of this strategy in robust model construction. This paper describes a multiple alignment scoring method and the results of our studies on HMM generation for distantly related protein sequences.

2 Material and Methods

2.1 Platforms and Software

All analyses were conducted on SUN Ultras (1/140 and 1/170) or SPARCstations (4, 5 or 10/514MP) running SunOS Release 5.5 or 5.6. Version 2.0 of Sequence Alignment and Modeling (SAM) was used for all studies. [Krogh, et al., 1994, Hughey and Krogh, 1996]

2.2 Data sets

Two types of sequence relationship distributions were used in these studies: 1) low-to-high sequence identity with high similarity; and 2) low identity, low similarity (LILS). Sequence identity is based on the number of common amino acid residues, while sequence similarity is based on the conservative substitution of amino acids.

In the studies presented here we tested various ranges (80-99%, 60-99%, 40-95% and 20-95%) of low-to-high sequence identity with high similarity relationships found among the RT proteins of the retroviruses. The LILS relationships ranged from 7-48% identity and included representatives from retrovirus, retrotransposon, retroposon and retrointron RT sequences. [McClure, 1993] The LILS data set includes an even distribution of RT sequences from the following groups: retroviruses (HT13, NVV0, SFV1, HERVC); *gypsy*-retrotransposons (GMG1, GM17, MDG1, MORG); *copia*-retrotransposons (CAT1, CMC1, CST4, C1095); retroposons (NDM0, NL13, NLOA, NTC0); and group II introns (ICD0, IAG0, ICS0, IPL0). GenBank accession number are L36905, M60610, X54482, M10976, M77661, X01472, X59545, Z27119, X53975, X02599, M94164, M22874, L19088, X60177, M62862, X98606, U41288, X71404, Z48620, with the exception of the *copia* agent which is from the Saccharomyces Genome Database.

2.3 Types of Models

Two types of models were tested. A *de novo* model is generated by training on each data set with internal sequence weighting to correct for sampling bias as provided by SAM. Then all twenty sequences are aligned to this model. A set of sub-class

models are generated when the sequences are differentially weighted as sub-classes based on the clustering of their pair-wise similarity scores. The LILS data set contains five sub-classes. The five sub-class models were generated by differentially weighting all sequences within one sub-class (4 sequences, 75% of total weight), relative to the other four sub-classes (16 sequences, 25% of total weight) during the training session. These weights are scaled to produce a sum of 20 which is equal to the sequence weight sum used in the *de novo* models. The end result is a set of sub-class models with amino acid probabilities at each node representing both the OSM and MIRs (figure 1). The four sequences belonging to each sub-class are aligned to their respective models. These alignments are then stacked together using an in-house program to create the final multiple alignment.

De novo and sub-class models were run: 1) with and without model surgery; and 2) with and without *a priori* knowledge of motif identity or location. Model surgery is a feature of the SAM that allows for the conversion of one state to another, or the addition or deletion of states after training based on number of sequences that invoke a particular state. *A priori* knowledge of motif identity and location is provided by the anchoring of the OSM within models, (figure 1).

A preliminary model, for use in the anchoring strategy, is created using the SAM program modelfromalign and the initial OSM alignment. The SAM modeling software also allows for designation of a number of special node types within the model. These special nodes are immune to model surgery. Two types of the special nodes are used in the studies presented here to anchor the OSM within a model. Type A nodes are invariant and cannot undergo further training. Type K nodes undergo transition training but not match or insert training. The core amino acid residues of the motif are assigned Type A nodes, while the amino and carboxyl residues of the motif are designated Type K. This designation allows for the transition training into and out of the Type A nodes representing the OSM.

In all of the initial models OSM anchoring is performed by designation of Type A and K nodes at the same positions in each model. Generic nodes are then added to represent the MIR equal to the largest number of amino acid residues present in each region in the sequence data set. The generic nodes are then trained by the SAM buildmodel program.

Each model type was trained on the data set with two different prior libraries: 1) the amino acid frequency of the training set, and 2) a 20-component Dirichlet mixture as provided in the SAM package.

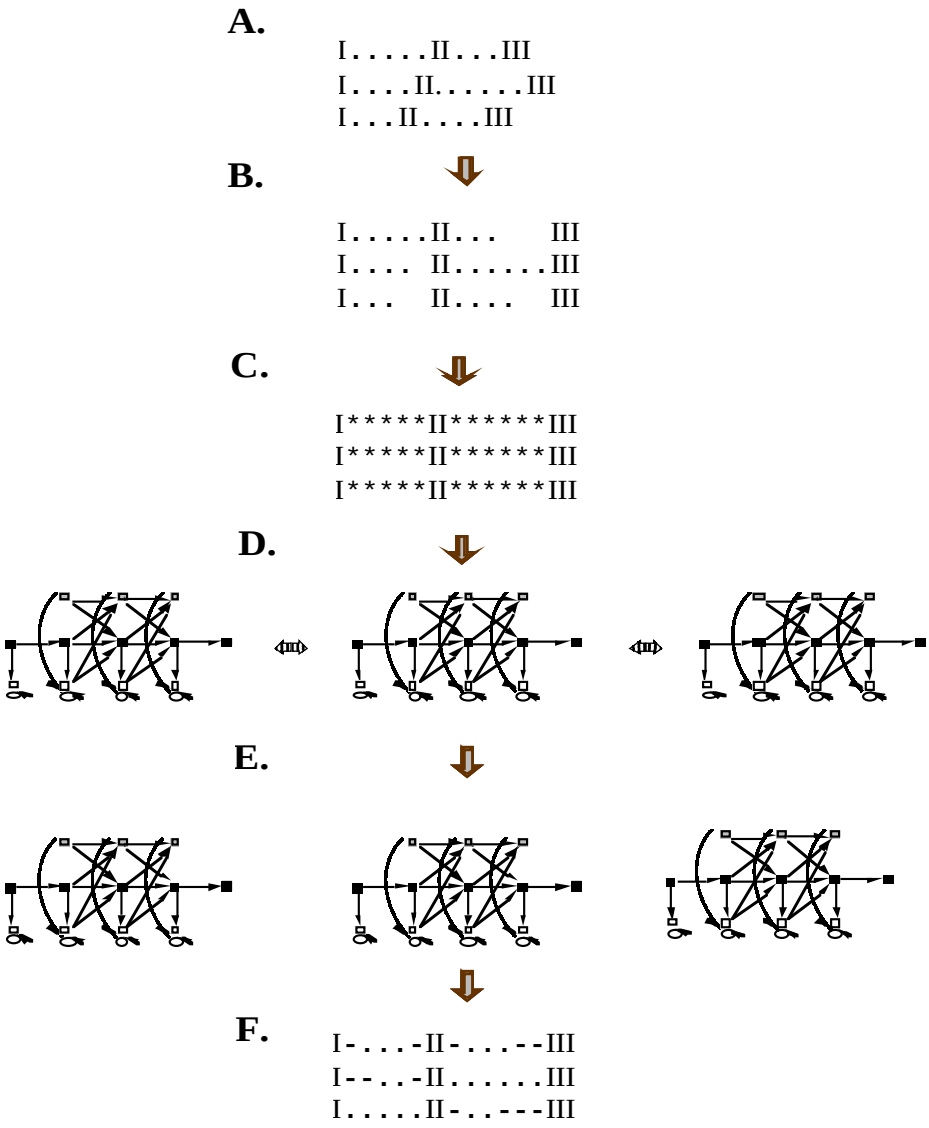

Figure 1. Schematic representation of OSM anchoring and sub-class modeling strategy.
A) Identification of the OSM in each sequence. B) The OSM is anchored by designation as special nodes at the same position in each model. C) The number of generic nodes added to all sub-class models between the OSM equals the largest number of residues in each of the MIRs. D) HMM modeling within and between sub-classes to align the MIRs. E) Sub-class models with amino acid probabilities at each node for the OSM and the MIRs aligned across sub-classes. F) Multiple alignment of the OSM and the MIRs. Roman numerals represent the OSM, and periods represent amino acid residues in the MIRs. Asterisks represent generic nodes. Dashes represent gaps.

2.4 Model parameter settings

All models were run with the same random seeds and at the default parameter settings except: Nmodels = 5, Nsurgery = 5, del_jump_conf = 50, match_jump_conf =50, ins_jump_conf = 50 and insconf = 100000. In the *de novo* models the internal_weight is 2. In the sub-classification models this parameter is set to zero so that our differential weighing is not modified. Inclusion of the external weights file is done using the sequence_weights parameter. The Dirichelet library is specified with the prior_library = recode2.20comp setting.

2.5 Methods for scoring HMM generated alignments

We have created an algorithm to score the multiple alignments generated by the test models. Algorithms that have used a column entropy measure to find conserved regions have proved successful. [Shenkin, et al., 1991] Since entropy calculations experience the smallest change in the OSM and the greatest change in the MIRs columns they are unsuitable for averaging over the entire alignment to generate a single score. Our algorithm is based on a column stability function similar to an entropy calculation that is averaged over the entire alignment length. All match, insert and delete states are included in the calculation. The stability measure algorithm is given by:

$$S = \left(\sum_{i=1}^{n} - (L_i / T)(\log(1.0 + c - (L_i / T))) \right) / n$$

Where S is the alignment score, n is the alignment length, L_i is the count of the largest group found at column i, T is the total number of sequences in the alignment, c is a constant currently set to 0.05, log is the logarithm base 2. The constant, c, can be any value greater than zero. It prevents the stability function from having a value of infinity with a full column count. It also allows for scaling of the stability values. At the current setting the column scores vary over the range from 0.003 for a 0% column count to 3.0 for a 100% column count. The current implementation of the algorithm produces three scores, M, M1, and M2, based on the largest group count of each column. The amino acid counts are currently based on three sets based on two levels of Dayhoff matrix conservative substitution: 1) the amino acid identities, (M); 2) ILMV, AG, ST, DE, NQ, C, FY, W, RK, H, P, (M1); and 3) ILMV, AGPST, DENQ, FYW, RKH, C, (M2). Each member of a group receives a count of one. This scoring method, that we call a stability measure, was designed to reflect the types of changes made by an expert in refining a multiple alignment. Expert refinements are introduced when obvious regions of identify or similarity are not detected by the alignment method or when alternative

positioning of insertions/deletions would either increase the similarity among the MIRs or minimize mutational events necessary to align one sequence to another. Our scoring method shows a positive correlation with the OSM count scoring used in our previous HMM construction studies. [unpublished data]

3 Results

All studies were conducted as described in the Material and Methods regarding parameter settings and prior libraries. In all studies the use of a 20-component Dirichlet mixture produced better alignments as assessed by the stability measure. [unpublished data and table 1] These results were expected due to the small size of the training set (20 sequences). [Brown, et al., 1993, Sjolander, et al., 1996] The OSM was found in the *de novo* HMM generated multiple alignments for the range tests of sequences with 80-99%, 60-99%, 40-95%, and 20-95% identity. The MIRs were also aligned in these alignments. [data not shown] No further analysis was conducted on these data.

table 1

	de novo, + surgery, -OSM anchor			*de novo*, + surgery, + OSM anchor		
	M	M1	M2	M	M1	M2
aa freq	0.052	0.109	0.150	0.073	0.129	0.175
D	0.050	0.099	0.138	0.090	0.163	0.221
	sub-class, + surgery, - OSM anchor			sub-class, + surgery, + OSM anchor		
	M	M1	M2	M	M1	M2
aa freq	0.052	0.108	0.150	0.030	0.064	0.094
D	0.049	0.097	0.133	0.030	0.062	0.092
	sub-class, - surgery, - OSM anchor			sub-class, - surgery, + OSM anchor		
	M	M1	M2	M	M1	M2
aa freq	0.052	0.108	0.150	0.089	0.153	0.202
D	0.049	0.097	0.133	0.106	0.192	0.245
expert refined alignment						
	M		M1		M2	
	0.127		0.216		0.274	

Definitions: aa freq = amino acid frequency of training set as calculated by SAM and D is a 20-component Dirichlet mixture provided in the SAM package. All other abbreviations are defined within the text

The LILS data (7-48% identity) provides a more challenging test of HMM construction. The generation of *de novo* models for the LILS data with and without OSM anchoring clearly indicates that by constraining the model in this manner more sequence relationship is found (table 1).

The second test of LILS data divided the training set into five sub-classes as described in the Material and Methods. Sub-class models were generated that allowed surgery, with and without OSM anchoring. These results indicate that allowing surgery in the MIRs defeats the keep node designation and shifts the location of various motifs within the OSM between sub-class models thereby lowering the stability measure on the final alignment (table 1).

The third study on the sub-classed LILS data did not allow surgery. As indicated in table 1 this approach provided the highest stability measure and reflects a better multiple alignment.

4 Conclusions and future studies

The motivation for these studies is the development of an automated method for the alignment of large numbers of highly divergent protein sequences that share common function and perhaps common ancestry. If the data used to train HMMs are not low identity and low similarity sequences then current HMM implementations work well. For LILS sequences, however, a more complex approach to HMM construction is necessary. Earlier work described the identification of the OSM as the first requirement for multiple alignment. [McClure, et al., 1994] We have devised and tested a strategy of HMM generation based on the anchoring of the OSM and sub-classification of the sequences. In these studies sub-models are built to represent the sub-classes. The sub-class alignments from these models are combined into a single multiple alignment. The goal of this approach is to maximize the alignment representation of the additional information contained in the MIRs.

Although in previous work we assessed the quality of HMM generated multiple alignments by the correct identification of the OSM, in these studies an independent scoring criterion, the stability measure, was designed to compare the multiple alignments. The stability measure incorporates the importance of the OSM and the MIRs in much the same way as a human expert.

By comparing the stability measures from the alignments generated by HMMs constructed under various constraints it is evident that OSM anchoring and sub-class modeling produces more informative multiple alignments than *de novo* models. This is due to increased alignment in the sub-class MIRs. The best multiple alignment generated in all these studies, however, was not as good as the alignment

refined by a human expert , (table 1) where alignment of the MIRs is maximized for all sequences .

Future studies will focus on improving the stability measure, further refinement of the OSM anchoring and sub-class model strategy to improve the alignment of the MIRs. Once we have determined a robust approach for modeling the MIRs, we hope to collaborate in the extension of current HMM implementations to incorporate this method.

Acknowledgments

This work was supported by a grant to M. A. M. from the NIH, AI 28309. We thank Angela Baldo and Julianna Hudak for useful scientific discussions, and Kevin Richter and Seanna Corro for help with manuscript preparation.

References

1. E. Domingo and J. J. Holland, "RNA virus mutations and fitness for survival" *Annu. Rev. Microbiol.* **51**, 151 (1997).
2. M. A. McClure, "Evolution of retroposons by acquisition or deletion of retrovirus-like genes" *Mol. Biol. Evol.* **8**, 835 (1991).
3. M. A. McClure, T. Vasi and W. M. Fitch, "Comparative analysis of multiple protein-sequence alignment methods" *Mole. Biol. Evol.* **11**, 571 (1994).
4. L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition" *Proceedings of the IEEE* **77**, 257 (1989).
5. P. Baldi, Y. Chauvin, T. Hunkapiller and M.A. McClure, "Hidden Markov models of biological primary sequence information" *Proc. Natl. Acad. Sci., USA* **91**, 1059 (1994).
6. A. Krogh, M. Brown, I. S. Mian, K. Sjolander and D. Haussler, "Hidden Markov models in computational biology: applications to protein modeling" *J. Mole. Biol.* **235**, 1501 (1994).
7. R. Hughey and A. Krogh, "Hidden Markov models for sequence analysis: extension and analysis of the basic method" *CABIOS* **12**, (1996).
8. A. McClure, in *Reverse Transcriptase*, "Evolutionary history of reverse transcriptase" (Cold Spring Harbor Laboratory Press, 1993).
9. S. Shenkin, B. Erman and L. D. Mastrandrea, "Information-theoretical entropy as a measure of sequence variability" *PROTEINS: Structure, Function, and Genetics* **11**, 297 (1991).
10. Sjolander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I. S. Mian and D. Haussler, "Dirichlet Mixtures: A method for improving detection of weak but significant protein sequence homology" *CABIOS* **12**, 327 (1996).