Low Identity, Low Similarity Protein Sequences: Independent Modeling of the Ordered-Series-of-Motifs and Motif-Intervening-Regions

Marcella A. McClure mars@parvati.lv-whi.nevada.edu
Julianna Hudak julie@parvati.lv-whi.nevada.edu
John Kowalski johnmk@parvati.lv-whi.nevada.edu

Department of Biological Sciences, UNLV, Las Vegas, NV 89129, USA

Abstract

We present a strategy for generating a multiple alignment from a hidden Markov model (HMM) for low identity, low similarity protein sequences. In this approach the ordered-series-of-motifs and the motif-intervening-regions are independently modeled. We also provide a measure of multiple alignment "goodness" called the stability function to compared one alignment to another. This strategy provides a more robust HMM representing highly divergent sequence data.

1 Introduction

RNA genomes (e.g., HIV, Ebola and Measles) can replicate and accumulate errors at a rapid rate. This ability creates a population called a quasi-species comprised of a consensus master genome sequence accompanied by a mutant cloud (Domingo and Holland 1997 [3]). These mutated RNA genomes provide a highly divergent set of co-linear genes encoding a variety of enzymatic and structural proteins. Many of the proposed relationships among the protein sequences encoded by these genes fail statistical criteria for homology (Schwartz and Dayhoff 1978 [20]; McClure 1992 [12]; Zanotto, Gibbs *et al.* 1996 [22]). 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 [11]).

The correct identification of the OSM among a set of protein sequences is the first step in multiple sequence alignment (McClure, Vasi *et al.* 1994 [16]). 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 MIR, however, can be constrained by selection pressures specific to sub-classes of the sequence set and often changes 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. The MIRs can define sub-class functional specificities and additional sub-class motifs. These regions contain information important in 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 low identity, low similarity protein sequences sufficient for maximal recovery of the OSM and MIRs remains a challenge in the field of bioinformatics.

The HMM approach to multiple sequence alignment (Baldi, Chauvin *et al.* 1994 [2]; Krogh, Brown *et al.* 1994 [8]; Eddy 1995 [5]; Hughey and Krogh 1996 [6]) provides a flexible method that can incorporate *a priori* knowledge into the model. We have demonstrated that anchoring an OSM in the same position within a set of subclass HMMs creates a series of models that generate better multiple alignments representing highly divergent sequences than a single model with or without OSM anchoring. In the course of these studies we also developed a basic stability measure to rank comparable multiple alignments (McClure and Kowalski 1999 [14]), although we are currently working on a refinement of this function. The studies presented here address two aspects of multiple alignment of highly divergent protein sequences: 1) a method for identifying the OSM; and 2) independent modeling of the MIRs.

2 Materials and Methods

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 multiple alignment studies (Krogh, Brown *et al.* 1994 [8]; Hughey and Krogh 1996 [6]).

2.1 Biological data

The protein family used in this study is the reverse transcriptase (RT), one of the two well-characterized domains of the RNA-dependent DNA-polymerase (RDDP). The RT domain is found in the amino portion of the RDDP encoded by different viruses (retroviruses, Hepadna-, Caulimo- and Badnaviruses), RNA-dependent transposable elements and other agents that inhabit a wide variety of Eubacterial and Eukaryotic hosts. The sequences in this study contain the well-defined OSM of the RT domain (Doolittle, Feng *et al.* 1989 [4]; Xiong and Eickbush 1990 [21]; McClure 1991 [11]) that has been confirmed to be of importance by X-ray crystal determination (Kohlstaedt, Wang *et al.* 1992 [7]).

The RT test sequences were obtained from GenBank, with the exception of one sequence from the Saccharomyces Genome Database. Two types of data sets were analyzed in the motif detection study, a 20 sequence data set and a 497 sequence data set. The 20 sequences were extracted from the 497 sequences using a program that generates pairwise similarity scores based on the Needleman-Wunsch algorithm (Needleman and Wunsch 1970 [17]), and CLUSTER, an in-house hierarchical clustering method. The pairwise sequence identity, based on the number of common amino acid residues, among this set of 20 sequences ranges from 7-48%. This range represents the average observed identity between the major groupings of RT sequences. The sequence similarity, based on the conservative substitution of amino acids, is also low. We refer to data of this type as low identity, low similarity (LILS) sequences. The 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); non-long terminal repeat retroposons (NDM0, NL13, NLOA, NTC0); and retrointrons (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.

2.2 Motif-identification programs

The motif-detection programs used in this study are MEME (Multiple Expectation Maximization for Motif Elicitation), PROBE and SAM. MEME and PROBE are local alignment methods that seek to locate and align an OSM without regard to the intervening regions. SAM is a global alignment method that attempts to align the entire length of a set of sequences. Initial studies included six local alignment methods for motif detection. This study only presents the results of two of these methods, MEME and PROBE, versus the SAM implementation of the HMM approach.

Both MEME and SAM methods locate motifs by estimating the parameters for a model that maximizes the likelihood of the data. MEME starts by breaking up the data into overlapping sequences of specified length (Bailey and Elkan 1994 [1]). The MM (Mixture Model) algorithm creates a finite mixture model of the new data set that consists of two components, the motifs and the motif-background probabilities. The Expectation Maximization algorithm estimates and maximizes the expected log likelihood value of the model parameters. The SAM program is a linear HMM that implements the Baum-Welch algorithm (Krogh, Brown *et al.* 1994 [8]; Hughey and Krogh 1996 [6]). The parameters estimated are the transition and observation probabilities. Once the model converges, a multiple alignment can be created and motifs detected.

The PROBE program implements the Smith-Waterman algorithm to perform transitive searches for finding regions of sequence similarity (Neuwald, Liu *et al.* 1997 [19]). The sequences collected from this search are purged to eliminate unequal representation of the data and then aligned co-linearly using the Gibbs sampling algorithm (Lawrence, Altschul *et al.* 1993 [9]; Neuwald, Liu *et al.* 1995 [18]). The Gibbs sampling algorithm starts with a random position for all of the sequences except one. The excluded sequence is aligned to the others. This process is reiterated until the information content score is maximized. After Gibbs sampling, a genetic algorithm is used to recombine a random alignment and select the best alignment produced. This alignment is used to search for more sequences, which are included in another iteration starting with the Gibbs sampling.

2.3 Strategy

2.3.1 Motif-identification

The best results for each method were obtained by performing parameter range studies with the LILS sequence data set. Initially, each program was executed using the default parameters. User-specified parameters were changed according to the description of their function and default values. A range of values for each parameter was tested to determine the effects on motif detection. The changes from the default parameters that produced significantly better results are included in Table 1. Parameter settings for the 497 sequence tests were derived from tests on the LILS data set.

Program performance is assessed by the correct identification of the amino acids of each motif (Fig. 1). Individual program scores consist of six values, one for each motif of the OSM. Each value is equal to the percentage of sequences in which the motifs are correctly identified.

2.3.2 Independent modeling of the MIRs

2.3.2.1 Types of MIR Models

A priori knowledge of the MIRs is provided by the identification of each motif of the OSM within each sequence. A multiple alignment in which only the OSM is modeled is used to extract the MIRs of each sequence. Each motif of the OSM is cut at the most conserved amino acid residue in an effort to provide a constraint on the edges of the MIRs. Each of the seven MIRs of the RT sequences is then independently modeled. There are two types of MIRs, internal and external. Internal MIRs are bordered by two motifs (more constrained) while external MIRs are not. Currently we do not distinguish between these two types of MIRs.

Two types of models representing each MIR were tested. A *de novo model* is generated by training each data set (20 sequences) with an internal sequence weighting to correct for sampling bias as provided by the SAM software. Determination of the number of sub-classes is based on the clustering of their pairwise similarity scores. The LILS data set contains five sub-classes. A *set of sub-class models* are generated by differential weighting of the sequences based on their inclusion/exclusion in each class. As determined in earlier studies, allowing model surgery can improve *de novo* modeling, but not sub-class models. As stated in the SAM manual, currently the feature performing surgery between sub-class models representing amino acids is not implemented.

De novo models are run with surgery and sub-class MIR models without surgery. Both types of models are run with and without OSM anchoring at MIR borders. Model surgery is a feature of the SAM software that allows the addition or deletion of states after training based on the number of sequences that invoke a particular state. The SAM software allows for designation of special node types within the model. The special nodes are immune to model surgery. Two types of special nodes are used in the studies presented

	I	II	III	IV	V	VI
HT13	pvkKa--	t-IDLkdaf	-LPQG-fk	qYMDDIll	shGLP-	kFLGqii
NVV0	ikkK---	tiLDIgday	-LPQG-wk	-YMDDIyi	qyGFM-	kWLGfel
SFV1	pvpKp--	ttLDLtngf	-LPQG-fl	aYVDDIyi	naGYV-	eFLGfni
HERVC	pvpKp--	tcLDLkdaf	-LPQR-fk	qYVDDLll	tvGIRc	cYLGfti
GMG1	mvrKa--	tkVDVraaf	-CPFG-la	aYLDDIli	--GLN-	kYLGfiv
GM17	v-pKkqd	ttIDLakgf	-MPFG-lk	vYLDDIiv	--NLK-	tFLG-hv
MDG1	lvpKksl	scLDLmsgf	-LPFG-lk	lYMDDLvv	--NLK-	tYLG-hk
MORG	vvrKk--	ttMDLqngf	-APFG-fk	lYMDDIiv	--GLK-	hFLG-hi
CAT1	lvdKpkd	eqMDVktaf	kSLYG-lk	lYVDDMli	-lSME-	rILGidi
CMC1	--tKrpe	hqMDVktaf	kAIYG-lk	lYVDDVvi	---KR-	hFIGiri
CST4	ftkKrng	t-LDInhaf	kALYG-lk	vYVDDCvi	inKLK-	dILGmdl
C1095	fnrKrdg	tqLDIssay	kSLYG-lk	lFVDDMil	itTLKk	dILGlei
NDM0	mihKt--	afLDIqqaf	gVPQGsvl	tYADDTav	nwNVR-	kYLGitl
NL13	lipKp--	s-IDAekaf	gTRQGcpl	lFADDMiv	vsGYK-	kYLGiql
NLOA	fipKa--	afLDIegaf	gCPQGgvl	gYADDIvi	evGLN-	kYLGvi-
NTC0	vlrKp--	amLDGrnay	gVRQGmvl	aYLDDVtv	alGIE-	rVLGagv
ICD0	eipKp--	vdIDIkgff	gTPQGgil	rYADDFki	rlDLDi	dFLGfkl
IAG0	fkkKt--	ieGDIksff	gVPQGgii	rYADDWlv	elKITl	-FLGvnl
ICS0	wipKp--	ldADIskcf	gTPQGgvi	rYADDFvi	emGLEl	nFLGfnv
IPL0	yipKs--	leADIrgff	gVPQGgpi	rYADDFvv	srGLVl	dFVGfnf

Figure 1: The six motifs of the RT OSM are indicated by roman numerals (I-VI). The bold and capitalized letters represent the core amino acids of each motif used to score the programs in this study. Dashes represent gaps in the alignment. Abbreviations on the left side bar are defined in Section 2 (materials and methods). The individual motifs of the OSM have varying levels of conservation. The order of conservation for the motifs, from high to low, is as follows: IV > II > VI > III > I or V. The OSM in the RT protein is well-characterized and these motifs are used to evaluate the performance of motif detection methods.

here to anchor the OSM within each MIR 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 as Type K. This designation allows for the transition training into and out of the Type A nodes representing the OSM.

In the sub-class MIR 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 use of sub-class weighting. The five sub-class models were generated by differentially weighting all sequences within one sub-class (75%), relative to the other four sub-classes (25% total) during the training session. The end result is a set of sub-class models with amino acid probabilities at each node representing both OSM and MIR that have been independently modeled.

The end result of MIR modeling produces separate models. Unfortunately the model concatenation subroutine of the MODIFYMODEL program available in the SAM software package does not currently work. An alternative approach is to align the appropriate subsection of each sequence to each MIR model. In-house software then stacks the MIR alignments and another program concatenates the sub-sections into a complete multiple alignment for scoring.

Each data set was used to train each model type 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.

2.3.2.2 Model parameter settings and alignment scoring

All models were run 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 (McClure and Raman 1995 [15]). In the *de novo* models the internal_weight = 2. In the sub-classification models this parameter is set to zero so that our differential weighing is not modified.

The multiple alignment scoring method used in the evaluation of alignments generated by the HMMs is designed to reflect the types of changes made by a human expert in refining a multiple alignment. Given that we cannot know all the possible mutations of the fast evolving MIRs, a parsimony approach is taken in the refinement. Changes 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 MIR 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.

The stability measure algorithm is given by:

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

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 0.03, 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 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: 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.

3 Results

3.1 Identification of the RT OSM

The best results from the 20 LILS sequences (a fairly smooth distribution representing the larger data set) and the 497 sequences (highly biased towards retroviruses) are presented in Table 1. For each motif, a score is reported in terms of the percentage of sequences in which the motif was correctly identified. All three methods detected the OSM for all 20 sequences in the LILS data set (Fig. 1) to some degree. Although the SAM method locates the OSM, it does not perform well in the recognition of individual motifs. All motifs, except motif II, are detected as subsets of correctly aligned motifs that are not aligned with the largest set of correctly identified motifs (indicated by an asterisk in Table 1). Scores for MEME and PROBE demonstrate that differences in program performance are not significant for the 20 LILS data set. The results of both MEME and PROBE coincide with the known information about motif conservation. The results for the LILS test clearly indicate that the local methods, MEME and PROBE, outperform the global method, SAM.

Program results for the 497 sequence data set are presented in Table 1. As expected, the accuracy of the SAM method in identifying motifs increases with the number of training sequences. Results improved significantly for the modeling and motif identification of the 497 sequence data set.

In the large, biased data set test, both MEME and PROBE eliminate sequences from the results, but for different reasons. MEME excludes sequences when it is unable to locate a motif in that sequence.

Table 1: Motif Scores and Parameter Options for RT Sequences.

SEQUENCE#	PROGRAM	I(1)	II(3)	III(4)	IV(5)	V(3)	VI(3)	PARAMETERS[a]
20	MEME	95	100	100	100	70	95	mod oops; nmotifs=10; distance=0.01
	PROBE	90	100	100	100	75	100	$S = 500$
	SAM	50*	75	40*	50*	45*	30*	internal_weight=2; FIMs 10,20,30,40,50
497	MEME	71*	97	88	98	71	87	mod oops; nmotifs=10; maxw=10; maxsize=180,000
	PROBE	86	99	100	100	97	87	$S = 500 + 5000$[b]
	SAM	38*	87*	90*	100	85*	69*	internal_weight=2

Roman numerals indicate motifs and values in parenthesis indicate number of amino acids scored for in each motif. Values in the columns indicate the percentage of sequences in which the motif was correctly identified. [a]The parameter column indicates the changes which gave the best results: mod oops = motif distribution equal to one occurrence per sequence; nmotifs = number of motifs to find; distance = Expectation Maximization convergence criterion; maxw = maximum motif width to be detected; maxsize = maximum data set size in characters; S = value at which to purge similar sequences; and FIMs = free insertion modules inserted at these positions; other SAM parameters were changed according to (McClure and Raman 1995 [15]; McClure 1996 [13]).
[b] Percentages for PROBE are from 72 of the 497 sequences due to the maximum limit of the purge (S) value. *Reported score is the highest percentage of correct motifs detected; lower percentages of motifs (not reported) are detected as subsets not correctly aligned to one another.

For the 497 sequence test, this reduces the number of sequences reported in the results. MEME also produces six different data sets, one for each motif, because individual motifs may easily be detected in one subset of sequences, but not another. This further reduces the available sequences to be used for alignment of the entire OSM. In MEME, the scores reported in Table 1 are the percentage of sequences correctly identified out of 497. Compared to the 20 LILS test, MEME had lowered performance for all six motifs. It should be noted that in order to get improved performance with MEME the user-specified number of motifs to be detected must be greater than the actual number of motifs.

The PROBE program excludes sequences when they are over-represented in the data set. This purged data set is then used to find the OSM. In this study, the maximum value for the purge parameter limits the reported sequences from 497 to 72. This results in OSM detection among a single data set consisting of the same 72 sequences. For PROBE, the scores are the percentages of the sequences correctly identified out of 72. PROBE showed a greater percentage of detection for motif V and a slight decrease of detection for the other five motifs.

Comparison of the results of the 20 versus 497 sequence test indicates that sequence similarity distribution influences program results of MEME but not PROBE. MEME scores higher with the unbiased set of 20 sequences because when a specific motif is over-represented in the data set the program will not recognize a divergent form of the motif. Thus, an entire sequence will be excluded because it is under-represented. In contrast, PROBE purges a biased data set by reducing redundant sequences or sequences that are too similar to each other. After it purges the sequences, the data is equally distributed and produces high scores regardless of bias in the input data set. However, increasing the purge value to include more similar sequences will reduce the scores slightly (Table 1).

3.2 Independent modeling of the MIRs

An earlier set of studies evaluated the effects of model surgery and OSM anchoring on *de novo* and sub-class modeling of the LILS RT sequences (McClure and Kowalski 1999 [14]). For purposes of comparison those data are reproduced here in Table 2. As described in 2.3.2.1 seven MIRs are present in the RT sequences as defined by the identification of the OSM. Two different model types, *de novo* and sub-class, are evaluated with and without surgery and motif anchoring of the MIR ends. All test models are run twice with the initial priors as the amino acid frequency count of the training sequences and with the 20-component Dirichlet mixture. The best model is the one with the highest stability score (calculated

Table 2: Results of modeling the entire sequence and the effects of OSM-anchoring and sub-classification on HMM construction.

	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

from Equation 1) regardless of which initializations priors generated the model. The data indicate that, in general, independent MIR modeling increases the alignment stability score when compared to entire sequence modeling (Tables 2 and 3). Significant score increases are observed for models in which the OSM has not been anchored. Although the models that include OSM anchoring also perform better, these scores are effected to a lesser degree. The best model ($S = 0.245$) from the earlier studies, compares favorably with two of the three best models of the new data presented here ($S = 0.234$ and 0.241). The stability score for the best model in the new study is 0.260.

The lack of a significant difference in the scores for *de novo* models with and without OSM anchoring is not due to lack of improvement in the alignment. The current implementation of our stability function does not distinguish between isolated column matches and the OSM. Refinement of this function to increase the column count for motifs is in process. As stated in the SAM manual, currently the feature performing surgery between sub-class models representing amino acids is not implemented.

The *de novo* model, with surgery and without OSM anchoring performs better than the sub-class model without anchoring or surgery due to the different ways sequences are weighted (Table 3). In the *de novo* model all sequences are weighted to eliminate sampling bias. In the sub-class model the closely related sequences carry more weight than more distant sequences. In contrast when OSM anchoring constrains the model the sub-class models produce better alignments than all other strategies.

4 Discussion

The multiple alignment problem of distantly related sequences is not a new problem in bioinformatics. A new approach to this problem is the use of HMMs that can incorporate *a priori* knowledge about

Table 3: Results of independent MIR modeling and the effects of OSM-anchoring and sub-classification on HMM.

	de novo, + surgery, -OSM anchor			*de novo*, + surgery, + OSM anchor		
	M	M1	M2	M	M1	M2
aa freq	0.066	0.122	0.158	0.103	0.176	0.226
D	0.103	0.189	0.241	0.105	0.181	0.234
	sub-class, - surgery, - OSM anchor			sub-class, - surgery, + OSM anchor		
	M	M1	M2	M	M1	M2
aa freq	0.093	0.166	0.207	0.114	0.196	0.252
D	0.093	0.164	0.205	0.117	0.202	0.260
	expert refined alignment					
	M		M1		M2	
	0.127		0.216		0.274	

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

the sequences. The first step in aligning such sequences is the identification of the OSM which defines membership in a specific protein family. A previous study of global and local methods revealed that global methods outperform local methods in identifying the OSM of four different protein families (McClure, Vasi *et al.* 1994 [16]). Another comparative study of HMM approaches concluded that HMMs were as good as or better at OSM detection than classical dynamic programming algorithms. Although HMMs display improved performance, they are not 100% accurate (McClure and Raman 1995 [15]; McClure 1996 [13]).

We have analyzed a variety of new motif detection algorithms using the four benchmark protein families, globins, kinase, aspartic acid protease, and the RH domain of the RDDP, (Hudak and McClure, manuscript in preparation). The data presented here on the RT domain is an extension of this work in the context of HMM generation. In the test of 20 LILS sequences both MEME and PROBE scores indicate that these methods have a high accuracy in motif identification. The most conserved motifs, IV > II > VI > III, had a high occurrence of detection. Motif I consists of a single residue and motif V is highly divergent (Fig. 1) and, therefore, they are the most difficult to correctly identify. Both MEME and PROBE performed better than SAM.

Once the OSM is identified, the second stage of multiple alignment of distantly related sequences is the alignment of the MIRs. In the data presented here the best approach for alignment of MIRs is the anchoring of the motifs and sub-class modeling. Anchoring the ends of the MIRs provides information on the OSM structure which acts as a constraint on the modeling. Sub-class modeling increases the retrieval of information within the MIR because similar sequences influence the model more than distant sequences.

The concept of distinguishing between the motifs common to a set of sequences and the intervening regions in multiple alignment strategies is not new (Martinez 1988 [10]). We have applied this concept in the context of HMM generation. In the course of the studies presented here, we have discovered that entire sequence HMM modeling may not be the best method for motif identification. We recommend the use of either MEME, an HMM approach, or PROBE, a combination of Gibbs sampling and a genetic algorithm, to initially find motifs. This information can then be incorporated *a priori* into an HMM for entire sequence modeling. It is clear from this analysis that an automated HMM approach that distinguishes between OSMs and MIRs would provide a better approximation of alignments that have been refined by human experts.

References

[1] Bailey, T.L. and Elkan, C., Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Second International Conference on Intelligent Systems for Molecular Biology*, Stanford, University, Stanford, CA, AAAI Press, 28–36. 1994.

[2] Baldi, P., Chauvin, Y., *et al.*, Hidden Markov models of biological primary sequence information, *Proc. Natl. Acad. Sci., USA*, 91:1059–1063, 1994.

[3] Domingo, E. and Holland, J.J., RNA virus mutations and fitness for survival, *Annu. Rev. Microbiol.*, 51:151–78, 1997.

[4] Doolittle, R.F., Feng, D.-F., *et al.*, Origins and evolutionary relationships of retroviruses, *Quart. Rev. Bio.*, 64:1–30, 1989.

[5] Eddy, S., Multiple alignment using hidden Markov models, *Third International Conference on Intelligent Systems for Molecular Biology*, Cambridge, England, AAAI Press, 114-120, 1995.

[6] Hughey, R. and Krogh, A., Hidden Markov models for sequence analysis: extension and analysis of the basic method, *CABIOS* 12:95–107, 1996.

[7] Kohlstaedt, L.A., Wang, J., *et al.*, Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor, *Science*, 256:1783–1790, 1992.

[8] Krogh, A., Brown, M., *et al.*, Hidden Markov models in computational biology: applications to protein modeling, *J. Mol. Biol.*, 235:1501–1531, 1994.

[9] Lawrence, C.E., Altschul, S.F., *et al.*, Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment, *Science*, 262(5131):208–214, 1993.

[10] Martinez, H.M., A flexible multiple sequence alignment program, *Nucl. Acids Res.*, 16(5):1683–1691, 1988.

[11] McClure, M.A., Evolution of retroposons by acquisition or deletion of retrovirus-like genes, *Mol. Biol. Evol.*, 8:835–856, 1991.

[12] McClure, M.A., Sequence analysis of eukaryotic retroid proteins, *Mathematical and Computer Modeling, An International Journal*, 16:121–136, 1992.

[13] McClure, M.A., Parameterization studies for the SAM and HMMER methods of hidden Markov model generation, *Fourth International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA, 155–164, 1996.

[14] McClure, M.A. and Kowalski, J., The effects of ordered-series-of-motifs anchoring and sub-class modeling on the generation of HMMs representing highly divergent protein sequences. *Proceedings of the Pacific Symposium on Biocomputing*, Hawaii, In Press, 1999.

[15] McClure, M.A. and Raman, R., Parameterization studies of Hidden Markov Models representing highly divergent protein sequences, *28th Annual Hawaii International Conference on System Sciences*, Hawaii, IEEE Computer Society Press, 184–194 1995.

[16] McClure, M.A., Vasi, T.K., *et al.*, Comparative analysis of multiple protein-sequence alignment methods, *Mol. Biol. Evol.*, 11(4):571–592, 1994.

[17] Needleman, S. B. and C. D. Wunsch (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48(3): 443–453.

[18] Neuwald, A.F., Liu, J.S., *et al.*,Gibbs motif sampling: detection of bacterial outer membrane protein repeats, *Protein Science*, 4(8):1618–1632, 1995.

[19] Neuwald, A.F., Liu, J.S., *et al.*, Extracting protein alignment models from the sequence database, *Nucleic Acids Research*, 25(9):1665–1677, 1997.

[20] Schwartz, R.M. and Dayhoff, M.O., Matrices for detecting distant relationships, In *Atlas of Protein Sequences*, Dayhoff, M.O., (ed.), Natl. Biomed. Res. Found, 353–358, 1978.

[21] Xiong, Y. and Eickbush, T.H., Origin and evolution of retroelements based upon their reverse transcriptase sequences, *The EMBO Journal*, 9(10):3353–3362, 1990.

[22] Zanotto, P., Gibbs, M.J., *et al.*, A reevaluation of the higher taxonomy of viruses based on RNA polymerases, *Journal of Virology*, 70(9):6083–6093, 1996.