Pacific Symposium on Biocomputing 4:138-149 (1999)

A COMPARATIVE ANALYSIS OF COMPUTATIONAL MOTIF-DETECTION METHODS

J. HUDAK, M.A. MCCLURE
Department of Biological Sciences, University of Nevada, Las Vegas, 89154-4004, USA

The detection of motifs within and among families of protein sequences can provide useful information regarding the function, structure and evolution of a protein. With the increasing number of computer programs available for motif detection, a comparative evaluation of the programs from a biological perspective is warranted. This study uses a set of 20 reverse transcriptase (RT) protein sequences to test and compare the ability of 7 different computational methods to locate the ordered-series-of-motifs that are well characterized in the RT sequences. The results provide insight to biologists as to the usage, value, and reliability of the numerous methods available.

1 Introduction

Early work in protein pattern recognition suggested that islands of amino acids may be conserved in the same order of a given protein family. [M.O. Dayhoff *et al.*, 1983] Today, a region of amino acids that is conserved throughout the evolution of a protein family is called a motif. Motifs can be present among protein sequences either as a set of unique motifs or as a set of repeated motifs. When motifs occur in a specific order among a set of sequences, they can be thought of as an ordered-series-of-motifs (OSM), [M.A. McClure, 1991] or protein signature. The designation of protein signature refers to the OSM that characterizes a particular family of proteins.

There are two aspects of motif detection worth clarifiying. The first is the initial recognition of a unique motif pattern, or OSM, that defines a protein family. The second is the use of known motifs to identify potential functions in uncharacterized sequences. We are interested in new computational methods for the initial inference of an OSM. Our approach to motif detection is an attempt to find the OSM among highly divergent sequences in order to provide insight into the function, structure and evolution of the protein family.

OSMs are selectively constrained throughout the evolution of a protein family as a result of their importance to function and structure. Thus, an OSM can be defined in more than one biologically meaningful way. A functional OSM can be described by the residues of a catalytic site, e.g., the Asp-Asp (DD) motif of the reverse transcriptase (RT) protein sequence. An OSM may also define structural patterns, e.g., α-helices or β-sheets. A functional OSM can be superimposed on structural domains; e.g., the RT OSM location within the fingers, palm and thumb domains of the RT (figure 1). [L.A. Kohlstaedt *et al.*, 1992] Regardless of how the

OSM is defined, function and structure is maintained only when all motifs of the OSM are present and in the appropriate order relative to one another .

In retroviruses, the RT constitutes one functional domain of the RNA-dependent DNA-polymerase (RDDP). The other domain is the ribonuclease-H (figure 1). Primary sequence analysis shows that all known RT sequences contain an ordered series of six characteristic motifs (figure 2). [M.A. McClure, 1993] The crystal structure of the RT reveals the location of the structural folds confirming the functional importance of the OSM. [L.A. Kohlstaedt *et al.*, 1992] 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. Since the OSM in the RT protein is well-characterized, the RT sequences can be used to evaluate the performance of motif detection methods.

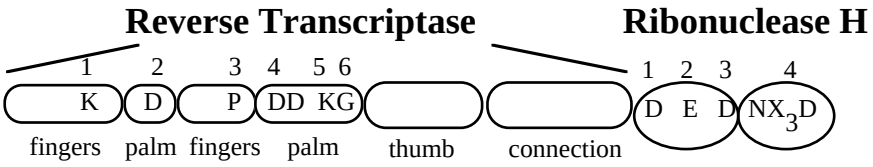

Figure 1. The RNA-dependent DNA polymerase (RDDP) is comprised of two functional domains, RT and RH. The most highly conserved residues of the OSM of the RT functional domain [M.A. McClure, 1993] are placed within the structural domains (fingers, palm, fingers, palm, thumb, and connection) identified by the HIV-1 RT crystal. [L.A. Kohlstaedt *et al.*, 1992] The most highly conserved residues of the ordered-series-of-motifs of the RH domain [M.A. McClure, 1991] are placed within the two RH structural domains based upon comparison of the HIV-I and *E. Coli* RH crystal structures. [J.F.I. Davies *et al.*, 1991]

With the increase in available sequence data, there has been an increase in computer programs created to define new motifs. Computational methods that attempt to identify an OSM without regard to the intervening regions are referred to as local alignment methods. Methods that attempt to align the entire length of a set of sequences are referred to as global alignment methods. A previous study of global and local methods revealed that global methods outperform local methods in identifying motifs. [M.A. McClure *et al.*, 1994] Another comparative study of global methods and HMM approaches concluded that HMMs were as good as or better at motif detection than classical dynamic programming methods. Although HMMs display improved performance, they are not 100% accurate. [M.A. McClure and R. Raman, 1995; M.A. McClure, 1996] With the increase in new computational methods for local alignment, a current comparative analysis is warranted. This study compares recently developed local alignment methods and the HMM approach.

From a biologist's perspective, choosing a computational motif-detection method is not simple, especially with the many different methods available. Once a method has been chosen, how does one know what parameters should be altered to produce optimal results? Comparative analyses of computational methods assist biologists in choosing and using the best method for their studies.

	I	II	III	IV	V	VI
HT13	pvk**K**a--	t-**IDL**kdaf	-**LPQG**-fk	q**YMDDI**ll	sh**GL**--	k**FLG**qii
NVV0	ikk**K**---	ti**LDI**gday	-**LPQG**-wk	-**YMDDI**yi	qy**GFM**-	k**WLG**fel
SFV1	pvp**K**p--	tt**LDL**tngf	-**LPQG**-fl	a**YVDDI**yi	na**GYV**v	e**FLG**fni
HERVC	pvp**K**p--	tc**LDL**kdaf	-**LPQR**-fk	q**YVDDL**ll	tv**GIR**c	c**YLG**fti
GMG1	mvr**K**a--	tk**VDV**raaf	-**CPFG**-la	a**YLDDI**li	--**GLN**-	k**YLG**fiv
GM17	v-p**K**kqd	tt**IDL**akgf	-**MPFG**-lk	v**YLDDI**iv	--**NLK**-	t**FLG**-hv
MDG1	lvp**K**ksl	sc**LDL**msgf	-**LPFG**-lk	l**YMDDL**vv	--**NLK**-	t**YLG**-hk
MORG	vvr**K**k--	tt**MDL**qngf	-**APFG**-fk	l**YMDDI**iv	--**GLK**-	h**FLG**-hi
CAT1	lvd**K**pkd	eq**MDV**ktaf	k**SLYG**-lk	l**YVDDM**li	--**EMK**-	r**ILG**idi
CMC1	tit**K**rpe	hq**MDV**ktaf	k**AIYG**-lk	l**YVDDV**vi	---**KR**-	h**FIG**iri
CST4	ftk**K**rng	t-**LDI**nhaf	k**ALYG**-lk	v**YVDDC**vi	in**KLK**-	d**ILG**mdl
C1095	fnr**K**rdg	tq**LDI**ssay	k**SLYG**-lk	l**FVDDM**il	it**TLK**k	d**ILG**lei
NDM0	mih**K**t--	af**LDI**qqaf	g**VPQG**svl	t**YADDT**av	ts**GL**--	k**YLG**itl
NL13	lip**K**p--	s-**IDA**ekaf	g**TRQG**cpl	l**FADDM**iv	vs**GYK**-	k**YLG**iql
NLOA	fip**K**a--	af**LDI**egaf	g**CPQG**gvl	g**YADDI**vi	ev**GLN**-	k**YLG**vi-
NTC0	vlr**K**p--	am**LDG**rnay	g**VRQG**mvl	a**YLDDV**tv	al**GIE**-	r**VLG**agv
ICD0	eip**K**p--	vd**IDI**k-gf	g**TPQG**gil	r**YADDF**ki	rl**DLD**i	d**FLG**fkl
IAG0	fkk**K**t--	ie**GDI**ks-f	g**VPQG**gii	r**YADDW**lv	el**KIT**l	-**FLG**vnl
ICS0	wip**K**p--	ld**ADI**sk-c	g**TPQG**gvi	r**YADDF**vi	em**GLE**l	n**FLG**fnv
IPL0	yip**K**s--	le**ADI**r-gf	g**VPQG**gpi	r**YADDF**vv	sr**GLV**l	d**FVG**fnf

Figure 2. The six motifs of the RT OSM are indicated by roman numerals (I-VI). [M.A. McClure, 1993] 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 materials and methods.

2. Materials and Methods

All analyses were performed on a Sun SPARCstation Ultra 1 running SUN OS 5.6.

2.1 Biological data

The RT test sequences were obtained from GenBank, with the exception of one sequence (C1095) from the Saccharomyces Genome Database. Initially, more than 500 RT sequences were retrieved from the databases. Using a program that generates pairwise similarity scores based on the Needleman-Wunsch algorithm, [S.B. Needleman and C.D. Wunsch, 1970] and CLUSTER, an in-house hierarchical clustering method, 20 representative RT sequences were selected from this collection. The pairwise sequence identity among the test set of sequences ranges from 7-48%. Based on the conservative substitution of amino acids, the sequence

similarity is also low. The dataset 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 numbers are L36905, M60610, X54482, M10976, M77661, X01472, X59545, Z27119, X53975, X02599, M94164, M22874, L19088, X60177, M62862, X98606, U41288, X71404, and Z48620.

2.2 Motif-identification programs

Seven computer programs were included in this study (table 1). With the exception of SAM, all of these programs are local alignment methods that are not search engines for motif databases. Although SAM is a global alignment method, it is included in this study because it was found to perform at least as well as global methods that are better than local methods. [M.A. McClure *et al.*, 1994] Brief descriptions of each program are provided below.

BLOCKMAKER, [S. Henikoff *et al.*, 1995] the downloaded version, implements the Motifj algorithm. [R.F. Smith and T.F. Smith, 1990] Motifj searches the sequences for conserved triplets of amino acids that are separated by a user-specified length. If the triplet is found in enough sequences, an alignment is created that maximizes the block score. From the best alignments, the triplets are merged and the alignment is extended to get the highest score for the blocks.

ITERALIGN uses the symmetric-iterative protocol. [L. Brocchieri and S. Karlin, 1998] It starts by aligning the sequences according to the significant segment pair alignment method. Improved sequences and, eventually, consensus sequences are generated until they converge. Blocks are derived from the alignment of the consensus sequences and are improved by displacement of individual sequences. The blocks are defined by a consensus residue and conservation index.

MATCHBOX implements a scanning algorithm. [E. Depiereux *et al.*, 1997] It begins the search using a 9-residue running window that moves across the sequences in search of a match. A match is based on the number of identical amino acids and the sum of the distances observed between matched residues. A database of matches/boxes is created and boxes are deleted based on their length or selected based on the residual length and gap cost ratio.

The PIMA (Pattern-Induced Multi-sequence Alignment) program starts by constructing a binary tree based on pairwise similarity scores. [R.F. Smith and T.F. Smith, 1992] The tree is reduced to one pattern by replacing nodes with a common pattern node that is generated by an alignment based on the Smith-Waterman (SW) algorithm. [T.F. Smith and M.S. Waterman, 1981] Common patterns are constructed from the alignment using amino acid class-covering hierarchy patterns.

The PROBE program implements the SW algorithm that performs transitive searches to find regions of sequence similarity. [A.F. Neuwald *et al.*, 1997] 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. [C.E. Lawrence *et al.*, 1993; A.F. Neuwald *et al.*, 1995] The Gibbs sampling algorithm starts at 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 randomly selected alignment and choose the best alignment produced. This alignment is used to search for more sequences, which are included in another iteration starting with the Gibbs sampling step, until no more new sequences are found.

Both MEME (Multiple Expectation Maximization for Motif Elicitation) and SAM (Sequence Alignment and Modeling) 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. [T.L. Bailey and C. Elkan, 1994] The MM (Mixture Model) algorithm creates a finite mixture model of the new dataset that consists of two components, the motifs and the motif-background probabilities. The EM (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. [A. Krogh *et al.*, 1994; R. Hughey and A. Krogh, 1996] The estimated parameters are the transition and observation probabilities. Once the model converges, a multiple alignment can be created and motifs detected.

Several programs are not included in this study for a variety of reasons. In a previous study, MACAW [G.D. Schuler *et al.*, 1991] and PRALIGN [M.S. Waterman and R. Jones, 1990] were found to give sub-optimal results. [M.A. McClure *et al.*, 1994] MOTIF [H.O. Smith *et al.*, 1990] was not included because it is only available for DOS and a modified version, Motifj, is implemented in the BLOCKMAKER program. The FILTER program was not suitable for this study due to a maximum sequence limit of 16. [M. Vingron and P. Argos, 1990; M. Vingron and P. Argos, 1991] PRATT was not included because detected motifs are based on PROSITE patterns. [I. Jonassen *et al.*, 1995; A. Brazma *et al.*, 1996] The EMOTIF program did not suit this study because it requires the input sequences to be aligned. [C.G. Nevill-Manning *et al.*, 1997] The TEIRESIAS program is not readily available. [I. Rigoutsos and A. Floratos, 1998] Initially, the GIBBS program was included. However, our analysis of GIBBS clearly indicates that the authors' most recent program, PROBE, performs better.

All programs were initially run at the default parameter settings to establish baseline results. Range studies for user-specified parameter options were conducted for all methods analyzed. Parameters were changed according to the description of their function and default values. A range of values for each parameter was chosen

to determine the effects on motif detection. The best results for each program were determined by a motif-scoring scheme.

2.3 Motif Scoring

Program performance was assessed by manually scoring the detected motifs. Individual program scores consist of six values, one for each motif of the OSM. The value for each motif is equal to the number of sequences correctly identified, with the highest score being the number of sequences (20) used to test the programs. The correct identification of a motif is based on the residues that represent the motifs (figure 2).

3. Results

The best results from these studies are presented in table 2. Of all the programs evaluated, ITERALIGN, MEME, PROBE, and SAM were the only ones that detected the entire OSM (figure 2). The highly conserved motif IV was the only pattern detected to some degree by all methods. The degree to which other motifs could be detected varied from method to method.

The webserver version of BLOCKMAKER implements both the Motifj and Gibbs sampling algorithms, without the option of changing parameters. The results for either algorithm are not any better than the downloaded version of Motifj with parameter changes. The best run of Motifj only detects the two most highly conserved motifs (figure 2; II and IV), with a high score of 19 for motif IV. The ITERALIGN program finds the entire OSM with motif VI (figure 2) having the highest occurrence of detection at 14 sequences. Parameter changes are not available for the webserver version of MATCHBOX. The only result obtained from this program is the detection of the most conserved motif IV in all 20 sequences. The highest scores (20) for MEME are for the two most conserved motifs (figure 2; II and IV). MEME also reports high scores for motifs I, III, and VI. PIMA detects all of the motifs except the highly divergent motif V. Motifs II and IV are detected in all 20 sequences while motif III is detected as two different unaligned subsets. The SAM method locates the entire OSM. All motifs, except motif II with a score of 15, are detected as unaligned subsets.

PROBE has the highest occurrence of detection for the entire OSM. These results were obtained after running the program several times under the default parameters. Differences in the results of these runs are due to different random seeds. The best random seed runs find the four most conserved motifs, II, III, IV, and VI, for all 20 sequences. Motif I, a single residue motif, was found in 18 out of 20 sequences. In two of the copia elements (figure 2; CAT1 and CMC1), the lysine residues were not correctly aligned. The highly divergent motif, V, was correctly

Table 1: Computational Motif-Detection Programs

PROGRAM	ALGORITHM[a]	MATRIX	INDEL PENALTY[c]	RUN TIME	USER SPECIFICATIONS[d] (# MOTIFS)	(WIDTH)	(# SEQUENCES)
BLOCKMAKER	Motifj	PAM 250	none	~1m	N	N	Ne
ITERALIGN	SI	PAM 250	C	~1h40m	N	Y	Y
MATCHBOX	Scanning	BLOSUM 62	none	~45m	N	N	Ni
MEME	MM/EM	PAM 250	none	~2m	Y	Y	Y
PIMA	SW	AACH[b]	I + E	~2m	N	N	Ni
PROBE	SW+G+GA	PAM 250	I + E	~2h30m	N	N	Y
SAM	BW	none	none	~2h20m	N	N	Ni

[a]Algorithms are: SI = Symmetric-Iterative protocol; MM = Mixture Model that uses (EM) Expectation Maximization; SW = Smith-Waterman; G = Gibbs Sampling; GA = Genetic Algorithm; and BW = Baum Welch. [b]AACH = Amino Acid Cluster Hierarchy (patgen, class 1; and class 2). [c]The indel penalties are: C = constant and I + E = initial + extension. [d]# MOTIFS = number of motifs to be detected; WIDTH = width of motifs to be detected; # SEQUENCES = number of sequences that contain the motif; N = user cannot specify; Ne=user cannot specify and program excludes sequences; Ni = user cannot specify, but program automatically includes all sequences; and Y = user can specify, but it is not required.

Table 2: Motif Scores and Parameter Options

PROGRAM	I(1)	II(3)	III(4)	IV(5)	V(3)	VI(3)	PARAMETERS
BLOCKMAKER	0	18	0	19	0	0	run type=1; sign=5; dist=5[a] (5-30)
ITERALIGN	10	9	8	13	12	14	ltw=0.99[b] (0.0-0.99)
MATCHBOX	0	0	0	20	0	0	default on webserver[c]
MEME	16	20	19	20	10	17	mod oops; nmotifs=10; maxw=10[d]
PIMA	18	20	8+12	20	0	15	default with class 2 matrix
PROBE	18	20	20	20	14	20	S=500[e]
SAM	10+2	15	8+5+3+2	10+3+2	9+2	6+2+2+2	iw=2; FIMs @ 10,20,30,40,50[f]

Roman numerals indicate motifs and values in parenthesis indicate number of amino acids scored for in each motif. Values in the columns indicate the number of sequences in which the motif was correctly identified. Some methods find correct matches in more than one subset of the data without correct alignment of these subsets to one another, indicated by more than one result per motif. The parameter column indicates the changes which gave the best results. Values in parenthesis in this column indicate the range over which a parameter was tested. [a]run type = 1 is non-iterative mode; sign = significance level; and dist = search width. [b]ltw = weight assigned to lower threshold hits. [c]no parameter changes available on the webserver. [d]mod oops = motif distribution equals one occurrence per sequence; nmotifs = number of motifs to find; maxw = maximum motif width to be detected. [e]S = level at which to purge similar sequences. [f]iw = internal_weight; FIMs = free insertion modules inserted at these positions; other parameters were changed according to (M.A. McClure and R. Raman, 1995, M.A. McClure, 1996).

identified in 14 out of 20 sequences. This motif was not correctly identified in any of the copia sequences and two non-long terminal repeat elements. Nonetheless, this study clearly indicates that the PROBE program outperforms all other methods (table 2).

Another strength of PROBE is that the results are reported as collinear blocks of motifs. Since collinearity is definitive of an OSM and block format is readily analyzed , this makes the result format of PROBE highly efficient. Other methods, such as BLOCKMAKER, MATCHBOX, MEME also display the results in a block format. However, MEME has a tendency to report motifs regardless of their position in the sequence. This is useful when looking for repetitive motifs throughout a set of sequences, but it does not maintain the collinearity of an OSM. Collinearity of BLOCKMAKER and MATCHBOX cannot be determined since the entire OSM was not detected. Methods, such as ITERALIGN, PIMA, and SAM display the results as an alignment of the data set. The alignments are collinear, but difficult to analyze. The motifs of the ITERALIGN alignment are difficult to score because the program allows gaps and insertions within the motif. PIMA reports motifs as a consensus sequence using 60 symbols that represent the different types of substitutions per position. This is difficult to analyze without a symbol legend and an alignment of the sequences to the consensus sequence. Since SAM is not meant for local alignment, it requires much effort to search the entire global alignment for the regions of aligned motifs.

4. Discussion and Future Studies

4.1 Discussion

The purpose of this study is to find the most reliable method of motif detection currently available. Motif-detection programs are sensitive to the degree of sequence similarity among the analyzed data. A program may be robust for analysis of similar sequences, but inadequate for a highly divergent set of sequences. Methods that are able to identify motifs among highly divergent sequences are more reliable than those methods that cannot.

While all programs analyzed were able to detect the most highly conserved motif IV, four of the methods (ITERALIGN, MEME, SAM, and PROBE) were able to detect the entire OSM. All other methods (BLOCKMAKER, MATCHBOX, and PIMA) were not able to identify motif V because it is one of the most divergent motifs. This indicates that although conserved motifs are easily detected, only the most robust methods will be able to detect an entire OSM that also contains divergent motifs. These results demonstrate that motif-detection programs are sensitive to the degree of sequence similarity.

Of all methods evaluated, PROBE performed the best at detecting the OSM in the highly divergent RT sequences. The PROBE program correctly located the four most conserved motifs and was able to detect the two divergent motifs with considerable accuracy. The error in detecting motif I for two sequences is surprising, because the two correct residues are only out of column register by 1 and 2 positions, respectively. PROBE is a robust method for detecting an OSM even without making any parameter changes. This is because it is designed to locate motifs as they are found in an OSM, collinearly among a set of sequences. In this study, PROBE detected more than the six collinear motifs of the OSM. This is not an inaccuracy of the method, but a display of PROBE's superior performance. The additional motifs detected are actually recognized sub-motifs in the RT sequences. [M.A. McClure, 1993] PROBE detects both motifs and sub-motifs without any specification from the user. This is useful when the number of motifs is not known. MEME, on the other hand, requires the number of motifs to be specified. MEME performance is improved when the specified number of motifs is greater than the actual number of motifs. This generates some sub-motif detection, but not as accurately as PROBE.

Although MEME has scores almost as high as PROBE, a recent analysis of both MEME and PROBE using a data set of 497 RT sequences demonstrated that PROBE is still able to outperform MEME. [submitted to GIW, McClure, Hudak and Kowalski, 1998] The data set used in the study contained an unequal distribution of sequence similarity which resulted in some sequences, or motifs, to be over-represented. MEME will get trapped in a local optima by recognizing the biased motif as the correct motif and considering any divergent form incorrect. This results in the exclusion of the entire sequence, thus reducing the score and producing biologically uninformative results. PROBE, however, handles a biased data set by eliminating redundant sequences or sequences that are too similar to each other. Purging of sequences produces an equally distributed data set representative of the entire 497 sequences from which it can detect informative motifs with a high score.

A recent comparison of several methods that are also included in this study (ITERALIGN, BLOCKMAKER, MEME, and PIMA) came to similar conclusions about program performance. [L. Brocchieri and S. Karlin, 1998] ITERALIGN and PIMA were able to find the entire OSM of the Rec-A sequences. MEME displayed better performance than BLOCKMAKER. Contrary to our experience with MATCHBOX (table 2), the program correctly identified 6 out of 7 Rec-A motifs. With the exception of MATCHBOX, program performance was comparable between the two studies even though our study used more divergent sequences with shorter motifs.

Our study has elucidated that PROBE is a superlative method currently available for the detection of an OSM.

4.2 Future Studies

Future studies will attempt to find an OSM among a larger group of highly divergent protein sequences that share analogous function. In addition to the RT domain sequences, this data set will be include sequences from the RNA-dependent RNA polymerases (RDRP) found in all other RNA viruses (e.g., HIV, Ebola, and Measles). In this case, some sequences of the data set cannot be statistically shown to share common ancestry. This raises the question of whether the observation of an OSM is due to common ancestry versus sequence convergence.

Whether or not common ancestry is responsible for the limited sequence similarity detected between the RT and RDRP sequences is an open question. Several studies suggest a common ancestry among all RNA-dependent polymerases. [P. Argos, 1988; O. Poch *et al.*, 1989; M. Delarue *et al.*, 1990] These studies were prompted by the detection of the highly conserved Asp-Asp motif in the RDRP of polio [G. Kamer and P. Argos, 1984] which is also found in retroviruses. Although the Asp-Asp motif is conserved among some RDRPs and the RT domain, there are only three additional residues found in common among these proteins, whose lengths vary from approximately 300 to 2000 amino acids. A recent reevaluation of the multiple alignments that suggested these relationships concludes that there is a lack of statistically significant signal remaining among the sequences to claim common ancestry. [P.M.d.A. Zanotto *et al.*, 1996]

A more robust motif-detection algorithm may aid in addressing the ancestry versus convergence question regarding RDRPs and the RT domain of RDDPs. Future studies will use the most reliable motif-detection method, as determined from this study, to locate a potential OSM shared among the RDRPs and the RT domain. Finding a reliable OSM would assist in creating separate hidden Markov models (HMMs) representing the sequences of both the RDRPs and the RT domain based on a new OSM-anchoring approach [see McClure and Kowalski, in these proceedings]. By comparing the protein sequences of one group to the model of the other, these HMMs can be used to evaluate the possibility of common ancestry between these sequences. If the probability is significant, then it would be worthwhile to construct an HMM representing both the RDRP and RT sequences. This approach could provide statistical evidence to either support or refute common ancestry among all RNA-dependent polymerases.

Acknowledgments

This work was supported by a grant to M.A.M. from the NIH, AI 28309. We thank John Kowalski for technical computer assistance, Seanna Corro, Micah Potter, Kevin Richter, and Steven de Belle for their invaluable assistance in discussion and manuscript critique.

References

1. M.O. Dayhoff *et al.*, in *Methods in Enzymology*, Ed. 1983).
2. M.A. McClure, *Mol. Biol. Evol.* **8**, 835 (1991).
3. L.A. Kohlstaedt *et al.*, *Science* **256**, 1783 (1992).
4. M.A. McClure, in *Cold Spring Harbor Symposia*, Ed. (Cold Spring Harbor Laboratory, Cold Harbor, NY, 1993).
5. J.F.I. Davies *et al.*, *Science* **252**, 88 (1991).
6. M.A. McClure *et al.*, *Mol. Biol. Evol.* **11(4)**, 571 (1994).
7. M.A. McClure and R. Raman, in *Proceedings of the 28th Annual Hawaii International Conference on System Sciences* , Ed. L. Hunterand B. Shriver (IEEE Computer Society Press, 1995).
8. M.A. McClure, in *Proceedings, Fourth International Conference on Intelligent Systems for Molecular Biology* , Ed. D.J. States et al. (AAAI Press, Menlo Park, CA, 1996).
9. S.B. Needleman and C.D. Wunsch, *J. Mol. Biol.* **48**, 443 (1970).
10. S. Henikoff *et al.*, *Gene-Combis, Gene* **163**, GC17 (1995).
11. R.F. Smith and T.F. Smith, *Proceedings of the National Academy of Science* **87**, 118 (1990).
12. L. Brocchieri and S. Karlin, *Journal of Molecular Biology* **276**, 249 (1998).
13. E. Depiereux *et al.*, *CABIOS* **13(3)**, 249 (1997).
14. R.F. Smith and T.F. Smith, *Protein Engineering* **5(1)**, 35 (1992).
15. T.F. Smith and M.S. Waterman, *J. Mol. Biol.* **147**, 195 (1981).
16. A.F. Neuwald *et al.*, *Nucleic Acids Research* **25(9)**, 1665 (1997).
17. C.E. Lawrence *et al.*, *Science* **262**, 208 (1993).
18. A.F. Neuwald *et al.*, *Protein Science* **4**, 1618 (1995).
19. T.L. Bailey and C. Elkan, in *Proceedings, Second International Conference on Intelligent Systems for Molecular Biology* , Ed. R. Altman et al. (AAAI Press, 1994).
20. A. Krogh *et al.*, *Journal of Molecular Biology* **235**, 1501 (1994).
21. R. Hughey and A. Krogh, *CABIOS* **12(2)**, 95 (1996).
22. G.D. Schuler *et al.*, *PROTEINS: Structure, Function, and Genetics* **9**, 180 (1991).
23. M.S. Waterman and R. Jones, *Methods Enzymol.* **183**, 221 (1990).
24. H.O. Smith *et al.*, *Proceedings of the National Academy of Science* **87**, 826 (1990).
25. M. Vingron and P. Argos, *Protein Engineering* **3(7)**, 565 (1990).
26. M. Vingron and P. Argos, *Journal of Molecular Biology* **218**, 33 (1991).
27. I. Jonassen *et al.*, *Protein Science* **4**, 1587 (1995).

28. A. Brazma *et al.*, in *Proceedings, Fourth International Conference on Intelligent Systems for Molecular Biology* , Ed. D.J. States et al. (AAAI Press, 1996).
29. C.G. Nevill-Manning *et al.*, in *Proceedings, Fifth International Conference on Intelligent Systems for Molecular Biology* , Ed. T. Gaasterland et al. (AAAI Press, 1997).
30. I. Rigoutsos and A. Floratos, *Bioinformatics* **14(1)**, 55 (1998).
31. P. Argos, *Nucl. Acids Res.* **16**, 9909 (1988).
32. O. Poch *et al.*, *The EMBO Journal* **8(12)**, 3867 (1989).
33. M. Delarue *et al.*, *Protein Engineering* **3**, 461 (1990).
34. G. Kamer and P. Argos, *Nucleic Acids Res.* **12**, 7269 (1984).
35. P.M.d.A. Zanotto *et al.*, *Journal of Virology* **70(9)**, 6083 (1996).