Comparative Analysis of Multiple Protein-Sequence Alignment Methods

Marcella A. McClure, Taha K. Vasi, and Walter M. Fitch
Department of Ecology and Evolutionary Biology, University of California, Irvine

We have analyzed a total of 12 different global and local multiple protein-sequence alignment methods. The purpose of this study is to evaluate each method's ability to correctly identify the ordered series of motifs found among all members of a given protein family. Four phylogenetically distributed sets of sequences from the hemoglobin, kinase, aspartic acid protease, and ribonuclease H protein families were used to test the methods. The performance of all 12 methods was affected by (1) the number of sequences in the test sets, (2) the degree of similarity among the sequences, and (3) the number of indels required to produce a multiple alignment. Global methods generally performed better than local methods in the detection of motif patterns.

Introduction

Comparison of primary sequence information is rapidly becoming the major source of data in the elucidation of the molecular mechanisms of replication and evolution of all organisms. There are basically three levels in the analysis of primary sequence information: (1) the search for homologues, (2) the multiple alignment of homologues, and (3) the phylogenetic reconstruction of the evolutionary history of homologues.

Many multiple sequence alignment programs and various scoring schemes have been developed to analyze potential relationships among sequences. Although a review (Myers 1991) and a comparison (Chan et al. 1992) of some methods from a computational perspective are available, there are no studies to date that evaluate these methods from a biologically informed perspective. The purpose of this study is to evaluate the ability of existing software to correctly identify the ordered series of motifs that are conserved throughout a given protein family.

There are two biological approaches to the multiple alignment of protein sequences: one attempts to align homologous (ancestrally related) features, while the other attempts to align functionally or spatially equivalent features of a protein family. While there is considerable overlap in the alignments produced by methods with these two goals, the intents are distinctly different.

Key words: sequence comparison, multiple alignment, protein family motifs.

Present address and address for correspondence and reprints: Marcella A. McClure, Department of Biological Sciences, University of Nevada, Las Vegas, Nevada 89154-4004.

Mol. Biol. Evol. 11(4):571–592. 1994.

0737-4038/94/1104-0002$02.00

Multiple alignment methods are often used without knowledge of the assumptions implicit in their operation. We will assess the major academically produced methods available, regardless of their intent, and indicate the assumptions implicit in each of the methods (table 1). Our basic premise is that, regardless of the final goal, a method that cannot find the functional motifs that are highly conserved throughout a given protein family has diminished value for detecting new biologically informative patterns.

The multiple protein-sequence alignment problem may be divided into the following two conceptual steps: (1) the initial inference of an ordered series of motifs defining the limits of a protein family and (2) detection of the ordered series of motifs in other proteins, thereby expanding the family. Many software packages, both academic and commercial, rely on the existence of previously defined protein families to provide the motifs of the family. How are such protein-family patterns initially determined? Among highly conserved sequences (>50% identity) it is very difficult to deduce which residues of a protein are necessary for function or structure, on the basis of multiple alignment of protein sequences alone. Laboratory experiments can provide clues as to which residues are critical for function and structure, but few generalizations can be made from such studies. Among distantly related proteins (<30% identical residues), however, conserved residues often indicate the essentially invariable regions of the protein that are necessary for function or structure. When multiple alignments of such data are derived, however, it soon becomes apparent that the currently available methods are not very satisfactory. Even with the utilization of the most sophisti-

Table 1
Multiple Alignment Methods

Method (Developer)	Algorithm	Matrix[a]	Indels	Limits[b]	Assumptions[c]	Features[d]	Data Type[e]
Global:							
AMULT (G. Barton)	NW	Any	C		Y, S	R, SE	P
ASSEMBLE (M. Vingron)	Dot matrix NW	Log odds	I+E		Y, S		P
CLUSTAL V (D. Higgens)	WL	Any	I+E			I	P, N
DFALIGN (D.-F. Feng)	NW	Log odds	C	UP	Y, E, O		P
GENALIGN[f] (H. Martinez)	CW, NW	UM	I+E			SE	P, N
MSA (S. Altschul)	CL	PAM250	I+E	ROS	N	B, FA	P
MULTAL (W. Taylor)	NW	UM, PAM250	C		S	AP, FA	P
MWT (J. Kececioglu)	maximum weight trace	Any	C	ROS	N		P
TULLA (S. Subbiah)	NW	Any	RGW	10 sequences	S	R, SE	P
Local:							
MACAW (G. Schuler)	SW	PAM250		DOS	Y	SE, FA, MD	P
PIMA (P. Smith)	SW	AACH	I+E		Y	MD	P
PRALIGN (M. Waterman)	CW	PAM250	I+E[g]		Y	MD, MC	P, N[h]

[a] The matrices are log odds and PAM250 (Dayhoff et al. 1978); UM = unitary matrix (Feng et al. 1985); and AACH = amino acid cluster hierarchy (Smith and Smith 1990).

[b] UP = unpublished parameters; ROS = easily runs out of computer space, thereby limited to six sequences; and DOS = runs only on a DOS system with Windows.

[c] Y or N = yes or no to the question Has homology been established?; S or E = multiple alignment is of structural or evolutionary intent; and O = input sequences must be in nearest-neighbor order, and a program is provided for this purpose.

[d] R = user-specified no. of iterations for refinement; SE = statistical evaluation is provided; I = interactive mode so that user may choose intermediate alignments; FA = specified region can be forced to align; B = correction for bias of overrepresentation of sequences; AP = alteration of parameters between iterations; MD = user-specified motif density; and MC = user-specified degree of motif conservation.

[e] P = protein; and N = nucleic acid.

[f] Licensed to IntelliGenetics.

[g] This indel penalty applies to CWs only.

[h] A separate program is available for nucleic acids.

cated software developed to date, refinement of such relationships still relies on the visual pattern-recognition skills of the human operator. The initial inference of the motifs defining a protein family by primary sequence analysis, therefore, requires the combination of multiple

subject to insertion, deletion, and duplication. There are two features of motifs that must be considered in their evaluation. The first, the motif density, is the percentage of the sequences in which a given motif is present. The second, the motif conservation, is the degree to which a

Table 2
Scores for Programs Tested Using Globins

Program and No. of Sequences Tested	Motif I (7 residues)	Motif II (5 residues)	Motif III (5 residues)	Motif IV (5 residues)	Motif V (3 residues)	Parameters/Comments[a]
	Global Methods					
AMULT:						
12	100	100	100	100	100	Single-order alignment; defaults except: indel = 8 (4–10) and iteration = 1 (1–4)
10	100	100	100	100	100	
6	100	100	100	100	100	
ASSEMBLE:						
12	100	92	100	100	100	Defaults except: FIL-SUM algorithm
10		Did not perform alignment, since filter produces empty plots[b]				FIL-LOG, I = 8 (8–12)
6	100	100	100	100	100	
CLUSTAL V:						
12	100	92	100	100	100	Defaults; parameters tweaked are: pairwise: indel (1–8) and *k*-tuple (1–2); multiple alignment: I (6–12) and E (2–10)
10	100	92	100	100	100	
6	100	92	100	100	100	
DFALIGN:						
12	100	100	100	100	100	Defaults
10	100	100	100	100	100	
6	100	100	100	100	100	
GENALIGN:						
12	92 (67, 25)[c]	100	100	83 (67, 17)[c]	92 (67, 25)	Defaults except: match weight = 2; NW
10	90 (60, 30)	90	90 (50, 40)	80 (60, 20)	90 (60, 30)	
6	83	100[c]	83 (50, 33)[c]	67 (2 × 33)	67 (2 × 33)	Defaults except: match weight = 1; NW

MULTAL:						
12	100	90	100	100	100	Matrix weight[d] = 0–5; cycles[e] = 12;
10	100	90	100	100	100	indel = 20; window size = 15–50;
6	100	90	100	100	100	cutoff score = 900–300; span[f] = 8–128[g]
TULLA:						
10	90	80	80	80	80	RGW = 2–4–6; median 2 or 4 (2–12)
6	83	83	83	67	83	RGW = 8 (4–12)
			Local Methods			
MACAW:						
12	75	92	75	67	67	Cutoff score = 30 (20–30); MD = 50%
10	70	80	70	60	60	(25%–50%); result list size = 100, for
6	100	67	100	67	67	all subsets; several overlapping blocks[h]
PIMA:						
12	100	100	100	100	100	E = 0.33; ML clusters[i]
10	100	100	100	100	100	
6	100	100	100	100	100	SB clusters[i]
PRALIGN:						
12	67	67 (33, 2 × 17)	75 (33, 25, 17)	67 (33, 2 × 17)	84 (67, 17)	Window size = 20 (10–40); word size
10	50 (30, 20)	60 (3 × 20)	60 (3 × 20)	20	50	= 3 (3–5); MC = 1 (0–2); indel
6	67 (2 × 33)	33	33	0	50	= 0; MD = 30% (20%–50%)

NOTE.—The score for each test is calculated as a percentage of the no. of sequences in each data set in which the motif was identified. Some methods find the correct matches in >1 subset of the data without being able to align these subsets to one another. In these cases, the total percent correct match is a combined score of the subsets (values in parentheses). Abbreviations are as in table 1.

[a] Deviations from default parameters are indicated by a dash for a single data set and by a bracket for two data sets or for new parameters used in all tests. The explored range of parameter values is indicated in parentheses.

[b] ASSEMBLE tends to produce only "correct" results or nothing.

[c] Has gaps in motif(s).

[d] Specifies the mix ratio between the identity matrix and the PAM250 (e.g., a weight of 2 indicates a 0.8 [identity matrix] + 0.2 [PAM250] mix).

[e] Specifies the no. of attempts the program makes to merge subalignments.

[f] Pairwise distance upper limit for the comparison of all sequences.

[g] MULTAL allows the user to change parameters for each cycle. Thus, the range shown in some of the parameters indicates the change of that parameter for each cycle.

[h] Creates several blocks for each cluster. One has to manually (with the help of the MACAW editor) merge them to get the percentages for each cluster.

[i] Creates alignments by using two types of clusters, maximal linkage (ML) clusters (Smith and Smith 1990) and sequence branching (SB) clusters (Smith and Smith 1992).

Table 3
Scores for Programs Tested Using Kinases

Program and No. of Sequences Tested	Motif I (6 residues)	Motif II (1 residue)	Motif III (1 residue)	Motif IV (9 residues)	Motif V (3 residues)	Motif VI (3 residues)	Motif VII (8 residues)	Motif VIII (1 residue)	Parameters/Comments
	Global Methods								
AMULT:									
12	100	83	92	100	100	100	100	100	Tree-based alignment
10	100	90	90	100	100	100	100	90	Single order alignment; iteration = 4 (1–4)
6	100	67	67	100	100	100	100	100	
ASSEMBLE:									
12	83	58 (33, 25)	83	100	100	100	100	100 (67, 33)	Defaults except: FIL-SUM algorithm.
10	90	30	0	100	100	100	100	70	FIL-LOG, I = 8 (8–12)
6	67	0	0	100	100	100	100	50	
CLUSTAL V:									
12	100	92	92 (50, 42)	100	100	100	100	100 (58, 42)	Defaults; parameters tweaked are: pairwise: indel (1–8) and *k*-tuple (1–2); multiple alignment: I (6–12) and E (2–10)
10	100	80 (50, 30)	80	100	100	100	100	90 (50, 40)	
6	100	83	67	100	100	100	100	100 (67, 33)	
DFALIGN:									
12	100	100	100	100	100	100	100	100	Begin weighting sequence 3 with value 2
10	100	100	100	100	100	100	100	100	Begin weighting sequence 2 with value 2
6	100	100	100	100	100	100	100	67	Begin weighting sequence 2 with value 2

GENALIGN:									
12	100[a]	75 (42, 33)	83	100	100	100	100 (2 × 50)	92 (67, 25)	Defaults except: NW; match weight = 1
10	80 (60, 20)	60 (40, 20)	80	100	100	100 (2 × 50)	100 (2 × 50)	90	
6	67	50	83 (50, 33)	100 (2 × 50)	100 (2 × 50)	100 (2 × 50)	100 (2 × 50)	83	
MULTAL:									
12	100	75 (58, 17)	83 (50, 33)	100	100	100 (58, 42)	100	100	Cycles = 14; window size = 15–140; cutoff score = 900–200; all others as in table 2[b]
10	100	80	50	100	100	100	100	100	
6	83	33	67	100	100	100	100	100	
TULLA:									
10	90[a]	60	80	100	100	90	90	90	RGW = 8–10–12, median 8
6	83[a]	83 (50, 33)	67	100	100	100	100	33	Defaults
				Local Methods					
MACAW:									
12	67	0	75	100	100	83	100	0	Cutoff score = 30 (20–30); MD = 50% (20%–50%); result list size = 100, for all subsets; several overlapping blocks[c]
10	70	0	50	100	100	90	90	0	
6	100	0	0	100	100	100	100	50	
PIMA:									
12	100	92	92	100	100	100	100	100	SB clusters[d]; E = 0.33 (0.2–1.75)
10	100	90	100	90	90	90	90	50 (30, 20)	SB clusters[d]
6	100	100	67	100	100	100	100	100	SB clusters[d]; E = 0.5 (0.2–1.75)
PRALIGN:									
12	100	84 (2 × 42)	50 (33, 17)	33	75 (42, 33)	75 (42, 33)	33	33	Window size = 20 (10–40); word size = 3 (3–5); MC = 1 (0–2) indel = 0; MD = 30% (20%–50%)
10	90	80 (30, 2 × 20)	20	40	70 (40, 30)	60 (2 × 30)	30	30	
6	67 (2 × 33)	67 (2 × 33)	0	0	67 (2 × 33)	67 (2 × 33)	67 (2 × 33)	33	

NOTE.—All designations and abbreviations are as in tables 1 and 2.

[a] See footnote "c" of table 2.

[b] See footnotes "d"–"g" of table 2.

[c] See footnote "h" of table 2.

[d] See footnote "i" of table 2.

Table 4
Scores for Programs Tested Using Proteases

Program and No. of Sequences Tested	Motif I (3 residues)	Motif II (5 residues)	Motif III (3 residues)	Parameters/Comments
		Global Methods		
AMULT:				
12	92	58	83	Tree-based alignment; SD ordering[a]
10	90	80 (50, 30)	70 (40, 30)	Single-order alignment; indel = 8 (4–10); iteration = 1 (1–4)
6	67	0	50	Tree-based alignment; SD ordering
ASSEMBLE:				
12				
10	Did not perform alignment, since filter produces empty plots[b]			
6				
CLUSTAL V:				
12	100	75 (50, 25)	50 (2 × 25)	Defaults; parameters tweaked are: pairwise: indel (1–8), *k*-tuple (1–2); multiple alignment: I (6–12), E (2–10)
10	100	70 (40, 30)	70 (30, 2 × 20)	
6	100	0	67	
DFALIGN:				
12	100	100 (70, 30)	100	Begin weighting sequence 3 with value 2
10	100	100 (70, 30)	100	Begin weighting sequence 2 with value 2
6	100	50	83	
GENALIGN:				
12	92	67 (42, 25)[c]	58 (25, 2 × 17)	Defaults except: match weight = 4, deletion weight = 2; NW
10	90 (70, 20)	50 (30, 20)[c]	80 (60, 20)[c]	Defaults except: match weight = 2; NW
6	67	33	0	
MULTAL:				
12	83	58 (33, 25)	75 (50, 25)	Cycles = 14; cutoff score = 900–200; all others as in table 2[d]
10	90 (50, 40)	70 (30, 2 × 20)	90 (50, 40)	
6	50	0	33	
TULLA:				
10	70	50 (30, 20)	70 (40, 30)	RGW = 2–4–6 median 4 (2–12)
6	83	33	0	RGW = 6–8–10 median 8 (2–12)
		Local Methods		
MACAW:				
12	100	25	67	Cutoff score = 20 (10–20); MD = 25%, 30%, 33% (20%–50%); result list size = 100, for all subsets; several overlapping blocks[e]
10	100	30	70	
6	100	0	33	
PIMA:				
12	100	42 (25, 17)	42 (25, 17)	SB clusters[f]
10	100	60 (40, 20)	70	SB clusters[f]; E = 0.33 (0.2–1.75)
6	100	0	33	SB clusters[f]
PRALIGN:				
12	67 (2 × 33)	34 (2 × 17)	67 (2 × 25, 17)	Window size = 20 (10–40); word size = 3 (3–5); MC = 1 (0–2); indel = 0; MD = 30% (20%–50%)
10	100 (40, 2 × 30)	30	70 (30, 2 × 20)	
6	100 (3 × 33)	0	30	

NOTE.—All designations and abbreviations are as in tables 1 and 2.
[a] SD ordering uses the standard deviation between sequence pairs to form an order.
[b] See footnote "b" of table 2.
[c] See footnote "c" of table 2.
[d] See footnotes "d"–"g" of table 2.
[e] See footnote "h" of table 2.
[f] See footnote "i" of table 2.

analysis. In addition to the hemoglobins, therefore, we have analyzed three such data sets: the kinase family, the aspartic acid protease family (both eukaryotic and viral), and the RH region of both the RNA-directed DNA polymerase (the reverse transcriptase) and the *Escherichia coli* RH enzyme.

of three motifs that contribute to the active site of the enzyme. The most prominent motif is three consecutive, conserved residues—aspartic acid, threonine, and glycine (single-letter code, "DTG") (fig. 3). It has been suggested that the aspartic acid proteases evolved through duplication of a single-domain prototype (Tang et al.

Table 5
Scores for Programs Tested Using RH

Program and No. of Sequences	Motif I	Motif II	Motif III	Motif IV

MULTAL:					
12	92 (75, 17)	92 (58, 2 × 17)	75 (50, 25)	83	Cycles = 14; cutoff score = 900–200; All others as in table 2[c]
10	100 (70, 30)	90	80 (60, 20)	70	
6	100	83	67	83	
TULLA:					
10	100[b]	50	40	80 (2 × 40)	Defaults except: RGW = 8–10–12 median 8
6	100	50	67	50	
	Local Methods				
MACAW:					
12	58	42	58	17	Cutoff score = 20 (10–20); MD = 25%, 30%, 33% (20%–50%); result list size = 100, for all subsets; several overlapping blocks[d]
10	80	70	70	40	
6	83	67	67	67	
PIMA:					
12	83	75	67 (33, 2 × 17)	92 (42, 33, 17)	ML clusters[e]; E = 0.2 (0.2–1.75); I = 5.5 (5–7)
10	100 (80, 20)	80	80 (40, 2 × 20)	90 (70, 20)	ML clusters[e]; E = 0.33 (0.2–1.75)
6	100	100	67	83 (50, 33)	
PRALIGN:					
12	75	67 (2 × 33)	50 (33, 17)	17	Window size = 20 (10–40); word size = 3 (3–5); MC = 1 (0–2); indel = 0; MD = 30% (20%–50%)
10	80	80 (60, 20)	40	20	
6	83	67 (2 × 33)	33	50	

NOTE.—All designations and abbreviations are as in tables 1 and 2.

[a] See footnote "b" of table 2.

[b] See footnote "c" of table 2.

[c] See footnotes "d"–"g" of table 2.

[d] See footnote "h" of table 2.

[e] See footnote "i" of table 2.

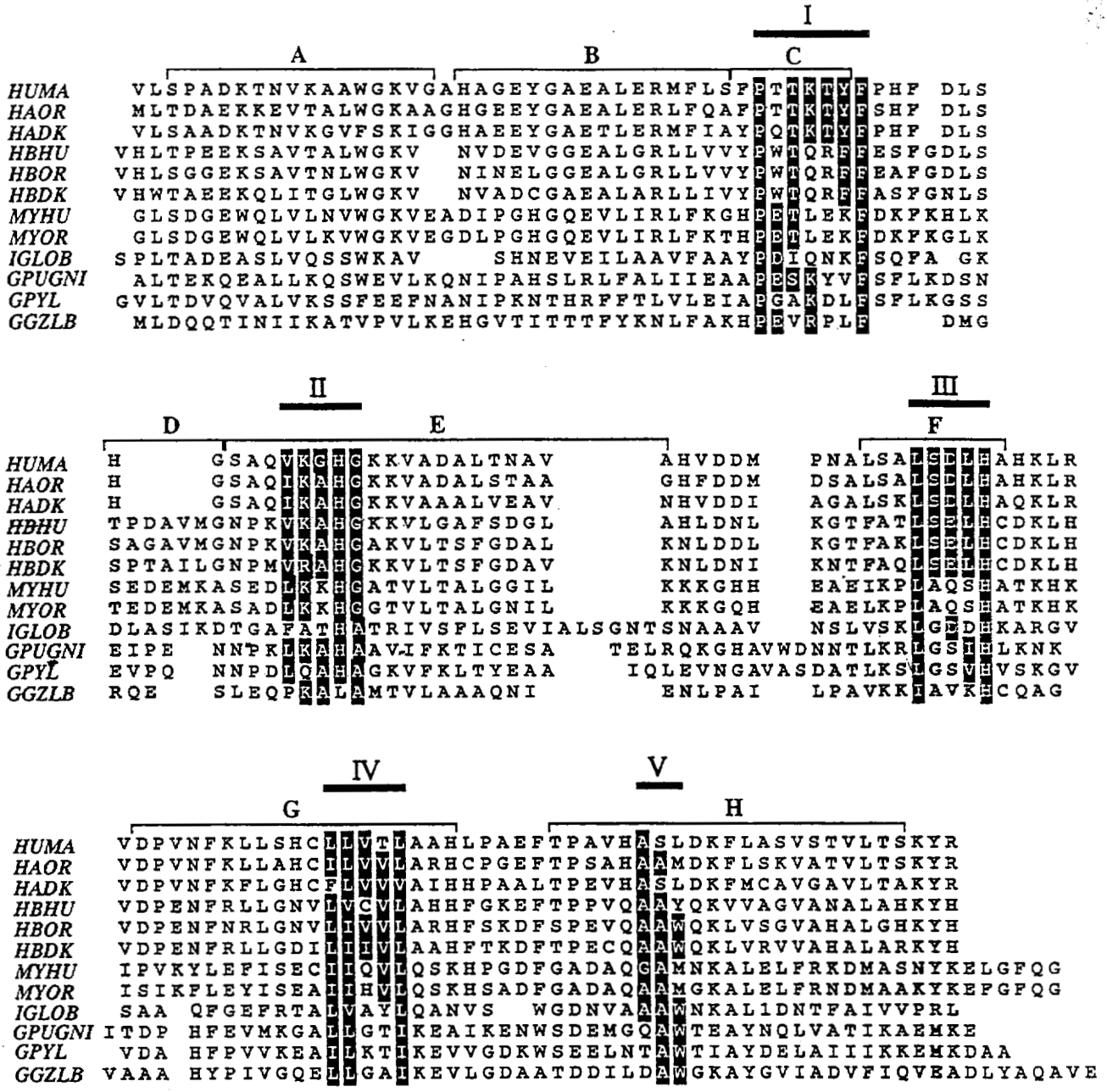
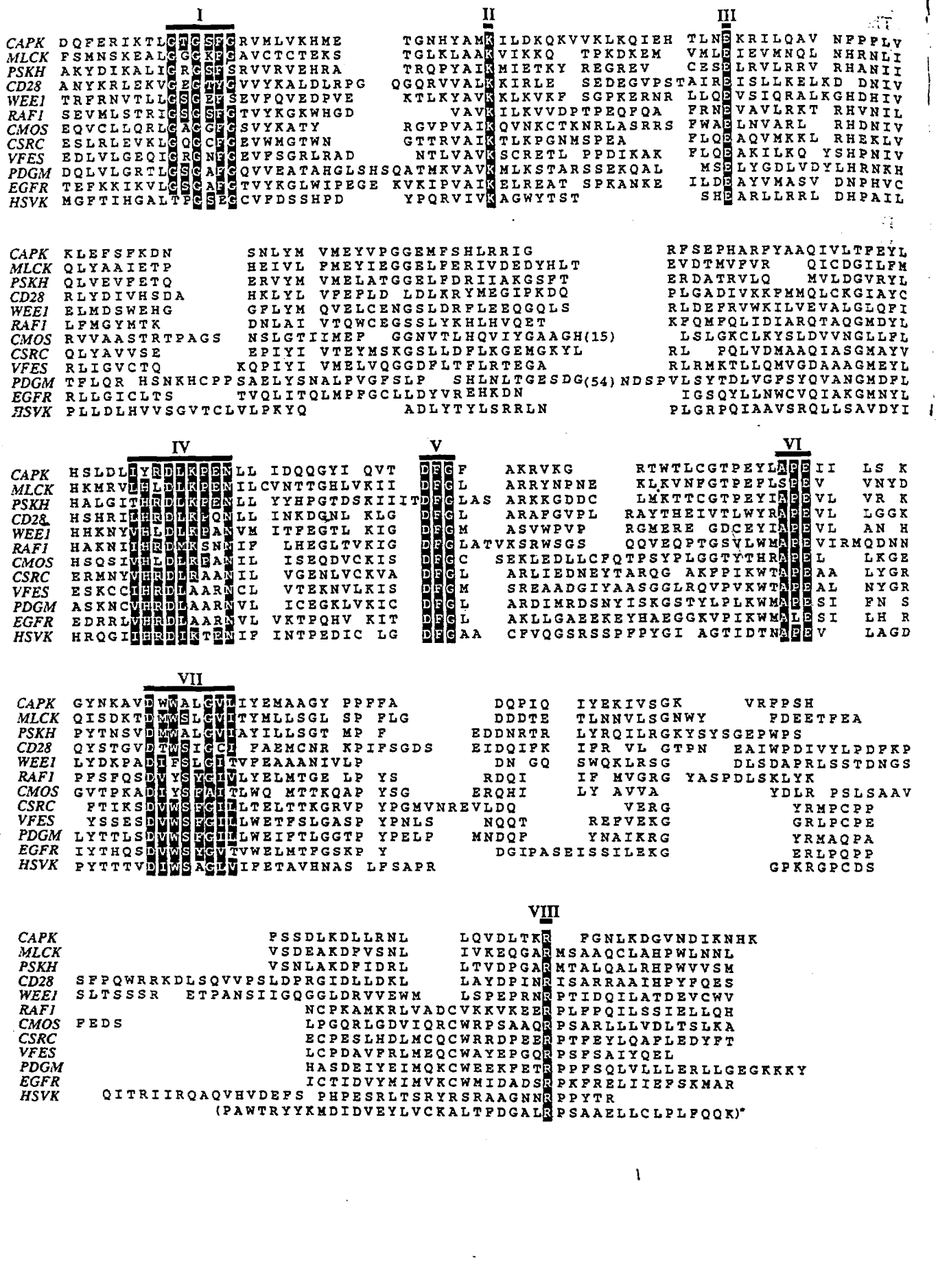


FIG. 1.—Multiple alignment of representative globin sequences. The five motifs scored for in the comparative analysis are indicated by blackened bars and the numerals I–V. Black/white reversals of columns within the motifs indicate the most conserved residues of the motifs and their conservative substitutions, based on the similarity scheme (F,Y), (M,L,I,V), (A,G), (T,S), (Q,N), (K,R), and (E,D). If the same number of matches occurs for more than one residue in a column, then one set is arbitrarily chosen for black/white reversal. The conserved helices of the globins are indicated by overlined regions and the letters A–H. The set of 12 sequences includes HAHU (human), HAOR (duckbill platypus), and HADK (duck) α-chain hemoglobins and HBHU (human), HBOR (duckbill platypus), and HBDK (duck) β-chain hemoglobins. MYHU (human) and MYOR (duckbill platypus) are myoglobins. The remaining hemoglobin sequences are IGLOB (insect, *Chironomus thummi*), GPYL (legume, yellow lupine), GPUGNI (nonlegume, swamp oak), and GGZLB (bacteria, *Vitreoscilla sp*). The two other test sets of globin sequences are subsets of these sequences; set 10 = set 12 without HAOR and HBOR, and set 6 is comprised of HAHU, HBHU, MYHU, IGLOB, GPYL, and GGZLB.

The sequences of the four protein families tested display a wide range of motif density, motif conservation, and indels. The globins are highly conserved with few indels, and the five motifs range in size from three to seven amino acids (fig. 1 and table 2). The kinase family has well-defined indel regions interspersed among eight highly conserved motifs, each of which varies from one to nine amino acid residues in size (fig. 2 and table 3).

The aspartic acid protease and RH sequences have the greatest range of motif density, motif conservation, and indels (figs. 2 and 3). The size of the three motifs of the protease is from three to five amino acid residues, and

sensus sequences to one another produces a progressive multiple alignment. In addition, GENALIGN allows the user to chose either the Needleman-Wunsch (NW) or consensus word (CW) algorithms (for definitions, see


I
II
III
IV
V
VI
VII
VIII
CAPK
MLCK
PSKH
CD28
WEE1
RAF1
CMOS
CSRC
VFES
PDGM
EGFR
HSVK

quences, to identify potential motifs. Only those motifs found in all pairwise alignments are coalesced into blocks that the user can then manipulate with the on-screen editor. The PIMA method begins with a pairwise analysis of all sequences, then constructs a tree on the basis of this order and derives a pattern at each node by using the progressive alignment approach (Smith and Smith 1990, 1992). This is continued in an iterative fashion until a root consensus pattern is achieved using the amino acid class hierarchy (see Scoring Matrices). PRALIGN is a method based on the CW approach (Waterman 1986; Waterman and Jones 1990). Words are found on the basis of user-specified word length (number of contiguous residues) and window length (number of consecutive residues to search within for a

of two sequences is created, and a dot is placed for matches. In the ASSEMBLE method the dot matrix is initially employed as a filter to identify and retain only those motifs that are conserved among a given set of sequences, prior to the use of dynamic programming. States and Boguski (1990) have written an elegant history and detailed description of the various biological applications of the dot matrix method.

Most of the methods compared here employ dynamic programming, which finds an optimal alignment for two sequences on the basis of various scoring schemes. The scoring scheme is usually based on a value for matches and replacements (see below) and on a penalty for indels (see below). The major shortcoming of this approach when applied to more than two sequences

does not allow the introduction of indels within a subsequence.

One global method (GENALIGN) and one local method (PRALIGN) are based on the CW approach to the multiple alignment problem (Karlin et al. 1983; Waterman 1986). It is assumed that the CWs defining a given protein family are unknown. All subsequences of a specific word size are then searched for within a given window among all the input sequences. Waterman and Jones (1990) have written a detailed description of the CW approach applied to both DNA and protein sequences.

Scoring Matrices

Various types of amino acid exchange matrices are available to assist in aligning protein sequences (Fitch and Margoliash 1967; Dayhoff et al. 1978; Feng et al. 1985; Taylor 1986; Rao 1987; Risler et al. 1988). Values for replacing one residue with another are based on physical/chemical similarities,

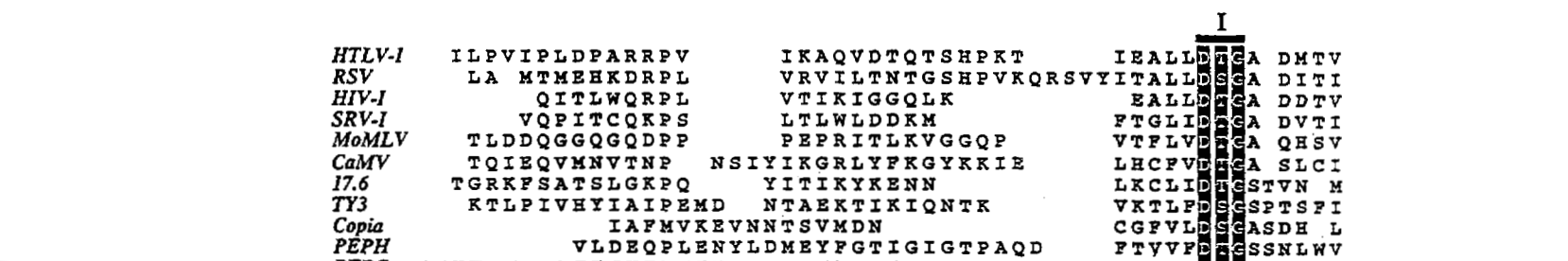
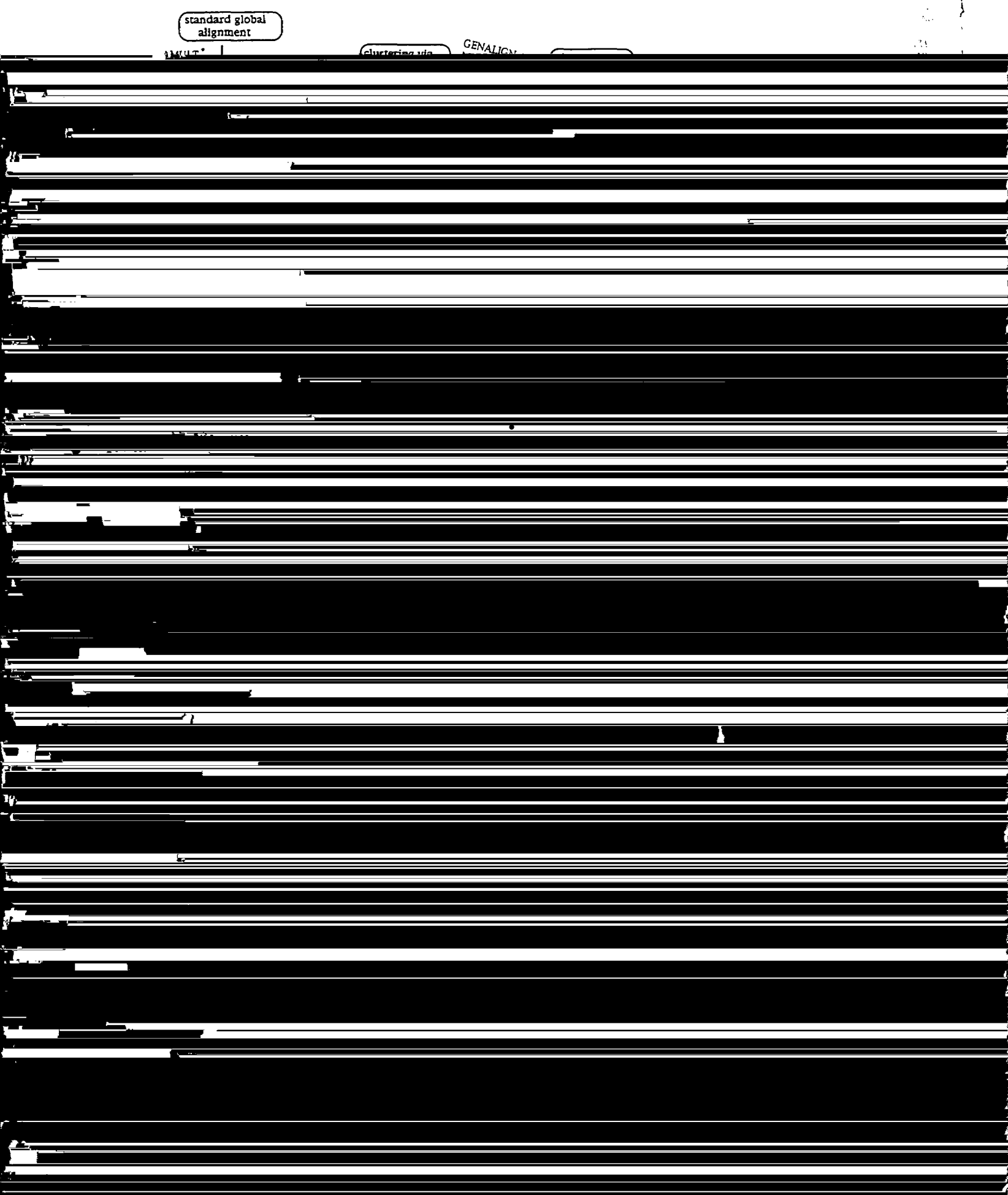


FIG. 4.—Multiple alignment of representative RH sequences. The four motifs scored for in the comparative analysis are indicated by blackened bars and the numerals I–IV. The retroid family RH sequences are from the retroviruses HTLV-II (human T-cell leukemia virus, type I) and HIV-II (human immunodeficiency virus, type II); the hepadnavirus HBV (human hepatitis B virus, ayw strain); the retroposon Ingi (*T. brucei*), and the group II mitochondrial plasmid Maup (Mauriceville, 1c strain) of *Neurospora crassa*. *Escherichia coli* is the ribonuclease H from *E. coli*. Other abbreviations are as in fig. 3. All other designations are as in figs. 1 and 2. The two other test sets of RH sequences are subsets of these sequences; set 10 = set 12 without HBV and Maup, and set 6 is comprised of PEPH, MoMLV, CaMV, COPIA, 17.6, and TY3.

ease of mutating one codon to another, and/or the observed frequency at which replacement occurs in closely related proteins. A widely accepted method

The amino acid class hierarchy is intrinsic to the PIMA method; therefore, this method cannot be evaluated with any other scoring scheme. This hierarchical


standard global alignment

produce results at all with our test sets. We attribute this to the space limitations of our computer (Kececioglu 1993). By using a set of six globins with >50% identity, however, MWT produces the correct alignment (unpublished observation). An implementation of the approximation algorithm for MWT that is space efficient is in progress (J. Kececioglu, personal communication). Future testing will determine whether either MSA or MWT can correctly identify motifs that define a protein family. These two methods will not be considered further.

Our comparative analysis indicates three distinct types of problems in multiple sequence alignment. The most significant problem encountered is the inability to merge subsets of sequences in which motifs have been correctly identified, to provide a single multiple alignment (tables 2–5). The global method GENALIGN and the local method PRALIGN exhibit this problem for all data sets to varying degrees, depending both on the number of sequences and on which specific sequences are analyzed (tables 2–5). In the kinase test, several other methods—ASSEMBLE, CLUSTAL V, MULTAL, TULLA, and PIMA—exhibit this problem to a minor degree. In this case the problem stems from the inability to recognize single-residue motifs that are common between subsets (table 3 and fig. 2).

Both the protease and RH data sets have some motifs that display low motif conservation (e.g., fig. 3, motif II, and fig. 4, motif IV). Most of the methods exhibit varying degrees of inability to merge correctly aligned subsets of sequences from these more distantly related data sets (tables 4 and 5). It should be noted that an additional weighting parameter was developed for DFALIGN (D.-F. Feng and R. F. Doolittle, personal communication) to specifically correct this type of error. This parameter allows the user to specify an additional weight (a value of 2 or 3 is sufficient) to be added to the score for each identical match beginning with a user-specified sequence. For example, in the kinase test set a weight of 2 is added for each identical residue common between sequences beginning with the third sequence. Use of this parameter is absolutely necessary to achieve the scores of tables 3–5 for the DFALIGN program. Extreme caution should be exercised in the manipulation of this parameter even by expert users (R. F. Doolittle, personal communication).

The second problem is the degree to which the number of sequences in the test set affects the ability to recognize motifs. Most methods perform better with larger data sets. In some cases, however, even though the accuracy of identifying motifs increases with the number of sequences, the inability to merge correct subsets of the data set is introduced into the multiple alignment (tables 3–5, comparing sets of 10 vs. 12).

The third problem, sensitivity to specific sequences in the data sets, appears to be a more general problem. One might think that the degree to which a method could identify motifs would not vary significantly as a function of addition or deletion of sister sequences to the data set, but only in the globin test is this problem negligible. Sensitivity to specific sequences is most consistently exhibited by the global methods GENALIGN and AMULT and by the local method PIMA, although all methods suffered to a degree from this problem (tables 2–5).

Discussion

Protein sequences with >50% amino acid residue identity can usually be unambiguously aligned by many of the multiple alignment methods currently available. Among protein sequences with <30% identity, it can be fairly straightforward to find the ordered series of motifs when the motifs are well conserved and when few indels have occurred (table 3 and fig. 2). It is difficult, however, to discern the ordered series of motifs that define a protein family and to obtain an adequate global multiple alignment that can be used in subsequent phylogenetic inference, if the motifs are not well conserved and if significant indels have occurred (tables 4 and 5 and figs. 3 and 4).

We have identified three specific problems that are exhibited to various degrees by all the methods tested. The first, the inability to produce a single multiple alignment, could be due to an indel penalty that is too high. This seems unlikely, since we have varied the indel penalties in most methods without alleviating this problem. The extra parameter of the DFALIGN method, which allows the user to increase the weight for matches as the distance between sequences increases, suggests that the inability to produce a single multiple alignment from subsets could be addressed as a matrix problem. Perhaps identical residues common among distantly related protein sequences should have a higher value, especially if they occur in small contiguous runs. The point, in the divergence of a family of protein sequences, at which such an increase in the values of identities should take precedence over more standard matrix scores needs to be investigated. Currently, subsets are merged by adjusting the placement of indels and appropriately reducing or increasing the number of indels to produce a single multiple alignment as a final manual refinement.

The second problem, the sensitivity to the number of sequences, and the third problem, which specific sequences are in the test set, are serious problems. The increase from 6 sequences to 10 sequences, by the addition of sister sequences to the test data sets, usually increases the ability of most methods to identify motifs. This increase, however, is accompanied by the intro-

duction of the inability to merge correct subsets. The addition of only two more sister sequences to the 10-sequence set, however, causes a decrease in identification of motifs. This effect is most significant for the protease and RH tests (tables 4 and 5). Why so many of the methods are sensitive to sequence number and specificity is an area that warrants further investigation on the part of the software developers. Such shortcomings should warn biologists that variation in data sampling could lead to erroneous conclusions regarding the ordered series of motifs defining a protein family, as well as the phylogenetic history of the gene, when these methods are used.

It is surprising that the global methods perform better than the local methods in the correct identification of the ordered series of motifs present in the four different data sets analyzed (tables 2–5). In addition, methods (global or local) based on the CW approach perform poorly compared with all other methods. In light of these results the biologist-user should exercise caution in the use of local methods or CW methods, either local or global, to infer functional motifs.

It is obvious that a method that can identify an ordered series of motifs, in which individual motifs can vary in both motif density and motif conservation, is just the first stage of obtaining a structural or evolutionarily meaningful multiple protein-sequence alignment. Once this is achieved, the intervening regions of the ordered series of motifs must be aligned. Such an alignment can then be used for phylogenetic reconstruction, for classification of additional sequences, and for determining significantly different subsequences among the sequences that will provide additional information about functional properties, e.g., substrate specificity.

We are interested in the development of multiple alignment approaches that are designed to reconstruct the evolutionary relationships between proteins. Such approaches must not only take into account sequence identity and conservative substitution based on mutational frequencies and physical and chemical similarities of amino acids, but must also be able to describe regions of indels and duplication that can be very useful as phylogenetic markers. Methods that only detect highly conserved motifs, while useful for inferring function, are insufficient for phylogenetic analysis. If all that is detected between proteins are the functionally or structurally constrained residues and if such regions form the basis of phylogenetic reconstruction, then one runs the risk of inferring an incorrect tree topology because of the increased likelihood of parallel or convergent substitutions; this problem can be mitigated by considering sequence information conserved between more closely related relatives.

The area of computational biology that encompasses both sequence-search and alignment algorithms has created a plethora of methods. In only a few instances have developers attempted to evaluate the multiple alignments produced by their methods by comparing them with experimentally determined structures (Barton and Sternberg 1987*a*, 1987*b*; Subbiah and Harrison 1989). The field is now sufficiently developed for adequate testing of methods on real sequence data. It is no longer sufficient that algorithm developers merely propose yet another approach to these problems. It is incumbent upon the software developers to specify the limits of new methods on the basis of an adequate sampling of known protein families. Likewise it is the obligation of the analytical biologist to provide well-controlled tests and to suggest further directions for the development of new methods for sequence analysis. Perhaps developers could use the test sequences described here to test new approaches versus older ones. We hope this study not only serves as a guide for multiple protein-sequence methods for biologists, but that it also provides an overview of the problem and a language with which to communicate with the mathematicians, statisticians, and computer scientists in the field. This analysis also provides the algorithm developers with a more informed perspective on the nature of the biological pattern recognition in primary sequences.

The ability to infer the ordered series of motifs that define a protein family is not trivial. While the parameter values utilized in the various methods analyzed in this study may serve as a guide for inferring motifs in other protein sequences, they should in no way be considered as *the parameters* that will always find the motifs. The state-of-the-art strategy for the initial inference of the motifs defining a protein family from primary sequence analysis still requires the combination of multiple alignment methods and human pattern-recognition skills.

Acknowledgments

We would like to thank all the developers who provided their source code and assistance. We are grateful to Mark Boguski, John Kececioglu, George Gutman, and Jacques Perrault for constructive criticisms on the manuscript. Support for M.A.M. and T.K.V. was provided by NIH grant AI 28309. Support for W.M.F. was provided by NSF grant DEB-9096152.

LITERATURE CITED

ALTSCHUL, S. F., R. J. CARROLL, and D. J. LIPMAN. 1989. Weights for data related by a tree. J. Mol. Biol. 207:647–653.

BARTON, G. J., and M. J. E. STERNBERG. 1987*a*. Evaluation and improvements in the automatic alignment of protein sequences. Protein Eng. 1:89–94.

———. 1987*b*. A strategy for the rapid multiple alignment of protein sequences confidence levels from tertiary structure comparisons. J. Mol. Biol. **198**:327–337.

BASHFORD, D., C. CHOTHIA, and A. M. LESK. 1987. Determinants of a protein fold unique features of the globin amino acid sequences. J. Mol. Evol. **196**:199–216.

CARRILLO, H., and D. LIPMAN. 1988. The multiple sequence alignment problem in biology. SIAM J. Appl. Math. **48**: 1073–1082.

CHAN, S. C., A. K. C. WONG, and D. K. Y. CHIU. 1992. A survey of multiple sequence comparison methods. Bull. Math. Biol. **54**:563–598.

DAVIES, J. F., Z. HOSTOMSKA, Z. HOSTOMSKY, S. R. JORDAN, and D. A. MATTHEWS. 1991. Crystal structure of the ribonuclease H domain of HIV-1 reverse transcriptase. Science **252**:88–95.

DAYHOFF, M. O., R. M. SCHWARTZ, and B. C. ORCUTT. 1978. A model of evolutionary change in proteins. Pp. 345–352 *in* M. O. DAYHOFF, ed. Atlas of protein sequence and structure. National Biomedical Research Foundation, Washington, D.C.

DOOLITTLE, R. F., D.-F. FENG, M. S. JOHNSON, and M. A. MCCLURE. 1989. Origins and evolutionary relationships of retroviruses. Q. Rev. Biol. **64**:1–30.

FENG, D.-F., and R. F. DOOLITTLE. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. **25**:351–360.

FENG, D.-F., M. S. JOHNSON, and R. F. DOOLITTLE. 1985. Aligning amino acid sequences: comparison of commonly used methods. J. Mol. Evol. **21**:112–125.

FITCH, W. M., and E. MARGOLIASH. 1967. Construction of phylogenetic trees. Science **155**:279–284.

GUSFIELD, D. 1993. Efficient methods for multiple sequence alignment with guaranteed error bounds. Bull. Math. Biol. **55**:141–154.

HANKS, S. K., and A. M. QUINN. 1991. Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members. Methods Enzymol. **200**:39–81.

HIGGINS, D. G., A. J. BLEASBY, and R. FUCHS. 1992. CLUSTAL V: improved software for multiple sequence alignment. Comput. Appl. Biosci. **8**:189–191.

JOHNSON, M. S., M. A. MCCLURE, D.-F. FENG, J. GRAY, and R. F. DOOLITTLE. 1986. Computer analysis of retroviral pol genes: assignment of enzymatic functions. Proc. Natl. Acad. Sci. USA **83**:7648–7652.

KARLIN, S., G. GHANDOUR, F. OST, S. TAVARE, and L. J. KORN. 1983. New approaches for computer analysis of nucleic acid sequences. Proc. Natl. Acad. Sci. USA **80**:5660–5664.

KATAYANAGI, K., M. MIYAGAWA, M. MATSUSHIMA, M. ISHIKAWA, S. KANAYA, M. IKEHARA, T. MATSUZAKI, and K. MORIKAWA. 1990. Three-dimensional structure of ribonuclease H from *E. coli*. Nature **347**:306–309.

KECECIOGLU, J. 1993. The maximum weight trace problem in multiple sequence alignment. Pp. 106–119 *in* A. APOSTOLICO, M. C. Z. GALIL, and U. MANBER, eds. The 4th symposium on combinatorial pattern matching. Springer, Berlin.

KNIGHTON, D. R., J. ZHENG, L. F. TEN EYCK, V. A. ASHFORD, N.-H. XUONG, S. S. TAYLOR, and J. M. SOWADSKI. 1991. Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. Science **254**: 407–414.

LIPMAN, D. J., S. F. ALTSCHUL, and J. D. KECECIOGLU. 1989. A tool for multiple sequence alignment. Proc. Natl. Acad. Sci. USA **86**:4412–4415.

MCCLURE, M. A. 1992. Sequence analysis of eukaryotic retroid proteins. Math. Comput. Modeling Int. J. **16**:121–136.

———. 1993. Evolutionary history of reverse transcriptase. Pp. 425–444 *in* A. M. SKALKA and S. P. GOFF, eds. Reverse transcriptase. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

MARTINEZ, H. M. 1988. A flexible multiple sequence alignment program. Nucleic Acids Res. **16**:1683–1691.

MILLER, M., M. JASKOLSKI, J. K. MOHANA RAO, J. LEIS, and A. WLODAWER. 1989. Crystal structure of a retroviral protease proves relationship to aspartic protease family. Nature **337**:576–579.

MYERS, E. W. 1991. An overview of sequence comparison algorithms in molecular biology. Tech. rep. TR 91-92. University of Arizona, Tucson.

NEEDLEMAN, S. B., and C. D. WUNSCH. 1970. A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol. **48**:443–453.

PEARL, L. H., and W. R. TAYLOR. 1987. A structural model for the retroviral proteases. Nature **329**:351–354.

PEVZNER, P. 1993. Multiple alignment, communication cost and graph matching. SIAM J. Appl. Math. **52**:1763–1779.

RAO, J. K. M. 1987. New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters. Int. J. Pept. Protein Res. **29**:276–281.

RISLER, J. L., M. O. DELORME, H. DELACROIX, and A. HENAUT. 1988. Amino acid substitutions in structurally related proteins: a pattern recognition approach: determination of a new and efficient scoring matrix. J. Mol. Biol. **204**:1019–1029.

SCHULER, G. D., S. F. ALTSCHUL, and D. J. LIPMAN. 1991. A workbench for multiple alignment construction and analysis. Proteins Structure Function Genet. **9**:180–190.

SMITH, R. F., and T. F. SMITH. 1990. Automatic generation of primary sequence patterns from sets of related protein sequences. Proc. Natl. Acad. Sci. USA **87**:118–122.

———. 1992. Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modeling. Protein Eng. **5**:35–41.

SMITH, T. F., and M. S. WATERMAN. 1981. Identification of common molecular subsequences. J. Mol. Biol. **147**:195–197.

STATES, D. J., and M. S. BOGUSKI. 1990. Similarity and homology. Pp. 89–157 *in* M. GRIBSKOV and J. DEVEREUX, eds. Sequence analysis primer. W. H. Freeman, New York.

SUBBIAH, S., and S. C. HARRISON. 1989. A method for multiple sequence alignment with gaps. J. Mol. Biol. **209**:539–548.

TANESE, N., and S. P. GOFF. 1988. Domain structure of the Moloney murine leukemia virus reverse transcriptase: mutational analysis and separate expression of the DNA polymerase and RNAase H activities Proc. Natl. Acad. Sci. USA 85:1777–1781.

TANG, J., M. N. G. JAMES, I.-N. HSU, J. JENKINS, and T. BLUNDELL. 1978. Structural evidence for gene duplication in the evolution of acid proteases. Nature 271:618–621.

TAYLOR, W. R. 1986. Identification of protein sequence homology by consensus template alignment. J. Mol. Biol. 188: 233–258.

———. 1987. Multiple sequence alignment by a pairwise algorithm. Comput. Appl. Biosci. 3:81–87.

———. 1988. A flexible method to align large numbers of biological sequences. J. Mol. Evol. 28:161–169.

VINGRON, M., and P. ARGOS. 1991. Motif recognition and alignment for many sequences by comparison of dotmatrices. J. Mol. Biol. 218:33–43.

WATERMAN, M. S. 1986. Multiple sequence alignment by consensus. Nucleic Acids Res. 14:9095–9102.

WATERMAN, M. S., and R. JONES. 1990. Consensus methods for DNA and protein sequence alignment. Methods Enzymol. 183:221–237.

WATERMAN, M. S., and M. D. PERLWITZ. 1984. Line geometries for sequence comparison. Bull. Math. Biol. 46:567–577.

WILBUR, W. J., and D. J. LIPMAN. 1982. Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. USA 80:726–730.

STANLEY A. SAWYER, reviewing editor

Received August 16, 1993

Accepted January 5, 1994