


# Finding Sequence Motifs in Groups of Functionally Related Proteins

HO Smith, TM Annau, and S Chandrasegaran

*PNAS* 1990;87;826-830
doi:10.1073/pnas.87.2.826

**This information is current as of November 2006.**

| | |
|---|---|
| | This article has been cited by other articles: www.pnas.org#otherarticles |
| **E-mail Alerts** | Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here. |
| **Rights & Permissions** | To reproduce this article in part (figures, tables) or in entirety, see: www.pnas.org/misc/rightperm.shtml |
| **Reprints** | To order reprints, see: www.pnas.org/misc/reprints.shtml |

Notes:

*Proc. Natl. Acad. Sci. USA*
Vol. 87, pp. 826–830, January 1990
Biochemistry

# Finding sequence motifs in groups of functionally related proteins

**(amino acid sequence/MOTIF program)**

HAMILTON O. SMITH*, THOMAS M. ANNAU*, AND SRINIVASAN CHANDRASEGARAN†

*Department of Molecular Biology and Genetics, School of Medicine, and †Department of Environmental Health Sciences, School of Hygiene and Public Health, The Johns Hopkins University, Baltimore, MD 21205

*Contributed by Hamilton O. Smith, October 27, 1989*

ABSTRACT **We have developed a method for rapidly finding patterns of conserved amino acid residues (motifs) in groups of functionally related proteins. All 3-amino acid patterns in a group of proteins of the type aa1 d1 aa2 d2 aa3, where d1 and d2 are distances that can be varied in a range up to 24 residues, are accumulated into an array. Segments of the proteins containing those patterns that occur most frequently are aligned on each other by a scoring method that obtains an average relatedness value for all the amino acids in each column of the aligned sequence block based on the Dayhoff relatedness odds matrix. The automated method successfully finds and displays nearly all of the sequence motifs that have been previously reported to occur in 33 reverse transcriptases, 18 DNA integrases, and 30 DNA methyltransferases.**

With the growth of the protein sequence data base in recent years, there have been a number of attempts to identify patterns of conserved amino acid residues in functionally related groups of proteins (1–11). Such sequence motifs can be found in some cases by single inspection, but more commonly they are found by examination of conserved residues in aligned groups of proteins. There are excellent computer programs for sequence alignment of two homologous proteins (12–14), but only recently have automated procedures for multiple sequence alignment been described (7, 15, 16). These generally utilize consecutive pairwise alignments, proceeding from the most homologous pair to the least. Lipman *et al.* (15) have implemented an algorithm capable of aligning up to eight sequences. While alignment methods are useful for discovering sequence motifs, they still require considerable effort when applied to large groups of proteins and they have the disadvantage that they cannot be applied to proteins bearing little homology.

We have developed an alternative method for identifying sequence motifs that depends on locating 3-amino acid patterns common to a majority of the sequences in a group. Our method works best on large groups of proteins (>10) and does not depend on the presence of global homologies. Thus its range of applicability is complementary to that of the alignment methods, which are most successful when applied to small numbers of homologous proteins.

## METHODS

**Definitions.** For purposes of this work we define the following terms (refer to Fig. 1 for examples). A *pattern* is any arrangement of two or more amino acids occurring in a protein. For example, A...Q.H..I is a 4-amino acid pattern in which the dots represent unspecified amino acids. A *motif* is a pattern that occurs frequently within a group of proteins. A motif is *stringent* if it occurs exactly in every member of a particular group of proteins, otherwise it is *degenerate*. A pattern occurring 7 times in a group of 10 proteins is a stringent motif with respect to the subgroup of 7 but is degenerate with respect to the group of 10. However, if the subpattern A...Q....I occurs exactly in the group of 10 proteins, then the subpattern forms a stringent motif for the whole group. A *block* is a solid array of aligned sequence segments containing a motif. Each row of a block is a solid (non-gapped) segment of one of the sequences from the group. Motifs, as we define them, are confined to sequence blocks. When multiple blocks are found, the individual blocks are generally separated by variable sequence distances from the other blocks.

```
AILGQSHMNI
AMPEQQRILI
AVGDQTRSAI
AVGEQLHAAI
AVANQHHWCI
AIKDQFHASI
AIDEQELASI
AMKEQAHRCI
ALIDQFHKCI
ALTNQGHSAI
```

FIG. 1. A sample motif block containing aligned sequence segments from 10 proteins. Blocks are solid; that is, they contain no gaps. The pattern A...Q....I, which occurs in all 10 sequences, is called a stringent motif, whereas the pattern A...Q.H..I, which occurs in only 7 of the 10 sequences, is called a degenerate motif. Motifs, as we narrowly define them, are simply patterns that occur frequently in a group of proteins and they are always confined to solid blocks. Whether or not a given motif has functional significance, rather than being merely a chance occurrence, must be determined by experiment.

**A Computer Program to Find Motifs.** We have developed a program (MOTIF) in the C language that rapidly finds 3-amino acid motifs in large groups of functionally related proteins. All 3-amino acid patterns of the type aa1 d1 aa2 d2 aa3 are accumulated into an array. The distances d1 and d2 can vary from 0 to 24 residues. The maximum range of variation for d1 and d2 is set at run time by means of a distance parameter. For example, if the distance range is set to 10 residues, the pattern array will contain $20 \times 10 \times 20 \times 10 \times 20 = 800{,}000$ elements. Only one pass through each sequence is required to tally the patterns; thus the time for the search is approximately proportional to the number of sequences and to the square of the distance parameter. Those patterns that occur a significant number of times and thus qualify as motifs are screened for redundancies. Duplicate 3-amino acid motifs, defined as motifs in which aa1 is at identical positions in the corresponding sequences of each motif, are screened out. These duplicates arise commonly. For example, a motif containing 5 amino acids will have six 3-amino acid motifs that start with the first amino acid and three 3-amino acid motifs that start with the second amino acid of the 5-amino acid motif. The seven duplicates are eliminated. In addition, since the program examines a min-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: cpu, central processing unit.

imum region of 30 residues surrounding each motif, only the highest-scoring motifs (see below), separated by <10 residues from each other, are retained. These two screening procedures serve to reduce the sometimes large number of related 3-amino acid motifs arising from a single larger motif.

We have developed a method of scoring motifs that provides a measure of the amount of surrounding homology and is useful for differentiating significant motifs from random background motifs (see below). A block containing a centrally located motif is divided into columns. To measure the degree of conservative substitution in each column, we use the PAM-250 scoring method developed by Dayhoff *et al.* (17), which gives a numerical score to each pair of amino acid residues in proportion to their frequency of substitution for one another by mutation during evolution. The PAM-250 scores of all pairs of residues in the column are totaled and averaged to form a mean column score. All positive mean column scores are then summed to form the motif block score. Motifs are displayed as a block, the top line of which summarizes the motif pattern by displaying a one-letter amino acid symbol whenever all residues in a column are identical, a star when the column score is ≥2, and a plus sign when the column score is 0 or 1. At the end of the block the 50% degenerate motif is displayed—i.e., the symbols of all residues that occur in a column with >50% frequency are displayed. As an option, the positions of motifs may be displayed on linear maps of the sequences.

MOTIF finds the best alignment of unmatched sequences not included in the block by sliding each unmatched sequence along the block, calculating a composite PAM-250 score at each position relative to the block, and choosing the highest scoring alignment. The composite score is a sum of products of the mean column score times the sliding sequence column scores. The latter are obtained by summing the PAM-250 scores of all pairs formed by a sliding sequence residue and a column residue from the block. Only positive column scores are used in the alignment score calculation. This rewards good residue fits and ignores poor fits. Matched sequences are also individually realigned against the block to select the best of two or more repeated patterns occurring in the same sequence and also to determine whether the original 3-amino acid match is the best match to the block overall.

Several parameters may be empirically set prior to compilation of the program. Currently the program uses the following default settings: the maximum number of sequences is 50, the maximum sequence length is 1000, the maximum distance between two residues in the 3-amino acid pattern is 24, the minimum and maximum block widths are 30 and 55, and the maximum total motifs and relevant motifs (after screening) are 100 and 50, respectively.

The number of matches selected as significant ($s$) for classifying a pattern as a motif is set during run time, as are the number of allowed internal pattern repeats ($r$) and the maximum value of the distance parameters d1 and d2 ($d$). A shorthand notation, [$s$, $r$, $d$], is used in the *Results* to indicate the parameter settings. Source code was compiled with Turbo C 2.0 (Borland). Computation (central processing unit, cpu) times were determined on a 12-MHz PC/AT-compatible (SIVA 286) computer without a math coprocessor.

**Sequences.** The DNA methylase sequences were obtained from G. Wilson (New England BioLabs). The DNA integrase sequences were supplied by R. Weisberg (National Institutes of Health) with the exception of HPint1, which was obtained from J. Scocca (The Johns Hopkins University). The reverse transcriptase sequences were from R. F. Doolittle (University of California, San Diego).

## RESULTS

**Estimating the Random Motif Background.** Our procedure searches for multiple occurrences of 3-amino acid patterns in a group of protein sequences. It is essential to know when a given multiple occurrence is statistically significant. A precise calculation of the expected frequency of random multiple occurrences of patterns in a given group of sequences is difficult; hence we will make some simplifying assumptions to obtain an estimate of the probability of such occurrences.

Consider $n$ random sequences each of length $l$ and with total length $L = n \cdot l$. Consider a single specific 3-amino acid pattern with the approximate mean occurrence of $m = L \cdot P_{aa1} \cdot P_{aa2} \cdot P_{aa3}$, where the $P$ values are the probabilities of each of the amino acids in the pattern. The probability of exactly $x$ occurrences of this pattern in the total length $L$ is the Poisson distribution term

$$P_x = (m^x e^{-m})/x!.$$

The probability of $<x$ occurrences is

$$P_{<x} = e^{-m}\left[1 + m + \frac{m^2}{2!} + \cdots + \frac{m^{(x-1)}}{(x-1)!}\right],$$

and the probability of $\geq x$ occurrences is

$$P_{\geq x} = 1 - P_{<x}.$$

But there are d1 · d2 · 8000 3-amino acid patterns. Therefore, the probability of $\geq x$ occurrences of at least one of the patterns is approximately

$$P = \text{d1} \cdot \text{d2} \cdot 8000 \cdot P_{\geq x},$$

where the approximation is good only if $P << 1$ and all the amino acids are of equal frequency.

Curve A in Fig. 2 shows the plot of the 50% expectation of $\geq x$ occurrences of at least one of the 800,000 3-amino acid patterns as a function of $n$ assuming d1 and d2 each independently range between 0 and 9 and all amino acids are of equal (0.05) probability. It can be seen that the background of chance multiple occurrences is quite significant relative to the number of sequences ($n$) when $n$ is <10, but as $n$ increases, the background becomes less significant relative to $n$. For example, when $n = 10$, there is a 50% chance of at least one pattern occurring 7 or more times, while with $n = 30$ the background level is 12. In real sequences, 3-amino acid patterns containing the most common amino acids will have the greatest mean occurrences and will account for most of

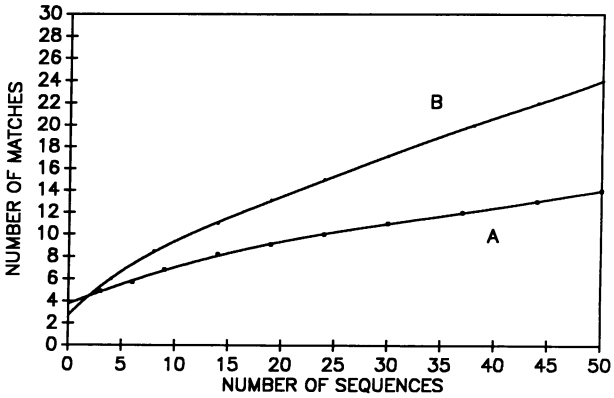



FIG. 2. Plot of the random background number of 3-amino acid pattern matches expected to occur in $n$ random sequences. The sequences are all 400 residues in length and the distance parameter is 10. In curve A the 20 different amino acids occur with equal frequency. In curve B, only the 2700 patterns composed of 3 amino acids with probabilities of occurrence of 0.08 are considered. For example, in curve B, 30 random sequences of 400 amino acids would have a 50% expectation of having at least one of the 2700 patterns matched 17 times.

the background. Curve B in Fig. 2 shows the level of background expected for just the 2700 patterns containing three particular amino acids with probabilities of 0.08 each. Clearly the background of chance multiple occurrences of such patterns is very significant.

An important feature of our program is its ability to filter out these random background motifs. By using relatively high significance levels, only patterns that occur at a frequency well above the random background level are obtained. In addition, random motifs are removed by setting the internal repeat parameter to 0 or to a low value. Random motif patterns tend to be repeated in one or more sequences, whereas real motif patterns tend to occur singly in each sequence of a group. Additional methods for distinguishing real from random motifs are covered in the *Discussion*.

**Sequence Motifs in DNA Methyltransferases.** Fifteen DNA-adenine methylase sequences were analyzed for sequence motifs. With the run time parameters set at [13,1,5], a single stringent motif, PPY (score 37), is identified and the corresponding 50% degenerate motif is D..Y.DPPY (Fig. 3). The same motif is repeatedly identified using distance parameters ranging from 1 to 24; however, computation time increases from about 2 sec to 214 sec over this range (see *Methods* for computer used), varying approximately as the square of the distance allowed between the 3 amino acids during the pattern search. With repeat and distance parameters fixed at 1 and 5, the same motif continues to be found as the significance factor is lowered. At settings of [9,1,5] the additional pattern F.G.G (score 29) appears, along with several apparently spurious motifs with scores generally <20 and consisting of the most common amino acids. With repeats set to 0, no motifs are detected until a setting of [9,0,5], when F.G.G and one apparently spurious motif (score 17) appear.

A routine is included in the program to randomly shuffle each sequence so that the background level of motifs can be measured empirically. With [9,0,5] parameter settings, no motifs are found after shuffling. With [8,0,5] settings, one or two motifs appear. With [9,1,5] settings, motifs range from three to six in number on independent shuffles. These motifs have scores <20 and contain the most common amino acids. The empirically determined level of background is close to that which we estimate from theoretical considerations (see Fig. 2).

Fourteen 5-methylcytosine methylases were examined. At parameter settings of [12,0,5], five motifs are found. These correspond to the five major motifs I, IV, VI, VIII, and X identified by Posfai *et al.* (7). At settings of [10,0,5], the additional minor motifs V, VII, and IX identified by those authors are found. No additional motifs are found at settings of [9,0,5] or at any other settings of the internal repeat or distance parameters. After the sequences are shuffled, no motifs are found at [9,0,5], one is found at [8,0,5], and 12 or 13 at [7,0,5].

The $N^4$-methylcytosine methylase M.*Pvu* II (18) was examined against either the fifteen adenine methylases or the fourteen 5-methylcytosine enzymes. With the 5-methylcytosine group, it showed high-scoring matches only to F.G.G. However, with the adenine methylase group, it scored well against both this motif and the DPPY motif.

The 30 pooled methylases were examined as a group. A single motif, F.G.G, is found in 23 proteins when the settings [23,1,5] are used. The 50% degenerate motif is LF.G.G .....G (Fig. 4). Partially matched residues are found in the remaining seven sequences. The F.G.G motif is found at lower significance-parameter settings down to 15. Below 15, the motifs present in the separate methylase groups begin to

Sequence: PPY in 13 proteins (1 duplicate), Score 37

```
   ++ *+** *PPY   + +
ENYSPKYNKAILNPPYLKIAAKGRERALLQ  260 (153)  192 (144)  mpsti
IGMVNRDDVVYCDPPYIGRHVDYFNSWGER  326 (194)  175 ( 96)  mecorv
MARADDASVVYCDPPYAPLSATANFYAYHT  316 (182)  195 (140)  mdam
IVDVRTGDFVYFDPPYIPLSETSAFTSYTH  328 (195)  189 (143)  mdpnii
LWKGGKFDFIVGNPPYVVRPSGYKNDNRIA  295 (120)  161 ( 49)  mcvibiii
APLEGQFDFVVGNPPYVRPELIPAPLLAEY  286 (121)  191 (  9)  mpaervii
WEPGEAFDLILGNPPYGIVGEASKYPIHVF  297 (106)  165 ( 51)  mtaqi
DVKILDGDFVYVDPPYLITVADYNKFWSED  328 (172)  195 ( 23)  mt4
TIPNESIDLIFADPPYFMQTEGKLLRTNGD  327 ( 37)  110 ( 32)  mhinfi
KMKPESMDMIFADPPYFLSNGGISNSGGQV  327 ( 37)  180 (  4)  mdpna
NAYAEKVNMIYIDPPYNTGKDGFVYNDDRK  331 (124)  173 ( 54)  mecop1
NAYAEKVKMIYIDPPYNTGKDGFVYNDDRK  326 (124)  173 ( 54)  mecop15
FSQLDQNDLVYCDPPYLITTGSYNDGNRGF  332 (549)  314 (219)  mfoki(219)
-----------------------------------------------------
IDLLKKSDIVVTNPPFSLFREYLDQLIKYD  249 (141)  170 ( 32)  mecori
LIPDKAVKIAFFDPQYRGVLDKMSYGNEGK  248 ( 38)  185 (  5)  mhhaii

.......D..Y.DPPY.............. Most common amino acids
```

FIG. 3. Alignment of 15 adenine methylase sequence segments about a common motif pattern. The first and third columns of numbers represent scores of the best and next best alignments. Numbers in parentheses are alignment positions of the first amino acid in the original 3-amino acid pattern. Numbers in parentheses following the sequence identifier name are positions of 3-amino acid pattern matches (internal repeats or original matches) that are not aligned. Sequences below the dotted line are unmatched to the primary 3-amino acid pattern. The best alignment is presented. MOTIF parameter settings were [13,1,5] and the cpu time to find the pattern was 11 sec.

Sequence: FGG in 23 proteins (1 duplicate), Score 29

```
    +  +*++F G G++  +++  + +
QYKSDKFNVLSLFCGAGGLDLGFELAGLEQ  244 ( 64)  129 ( 58)  mbspri
DFRTDKINVLSLFSGCGGLDLGFELAGLAA  244 ( 65)  144 ( 33)  mbsuri
......MNIIDLFAGCGGFSHGFKMAGYNS 247 (  7)   65 (  4)  mddei
HHPDYAFRFIDLFAGIGGIRKGFETIGGQC  251 (116)  140 ( 18)  mecorii
......MNLISLFSGAGGLDLGFQKAGFRI  242 (  7)   77 (  6)  mhaeiii
DKQLTGLRFIDLFAGLGGFRLALESCGAEC  248 ( 18)  132 ( 14)  mhhai
PNLREKFTFIDLFAGIGGFRIAMQNLGGKC  248 ( 38)  136 ( 32)  mhpaii
DAYSSDFKFIDLFSGIGGIRQSFEVNGGKC  250 (111)  132 ( 59)  mmspi
PTTYNPMKIISLFSGCGGLDLGFEKAGFEI  245 ( 18)  100 (  9)  mngopii(143)
...LSKLRVMSLFSGIGAFEAALRNIGVDY  242 ( 10)   71 (  4)  mphi3t
...LSKLRVMSLFSGIGAFEAALRNIGVEY  242 ( 10)   71 (  4)  mrho11s
...LGKLRVMSLFSGIGAFEAALRNIGVGY  240 ( 10)   71 (  4)  mspr
SEPKNKPKALSFFSGAMGLDLGIEQAGFET  197 ( 81)  129 ( 80)  msini(343)
HHPHYAFRFIDLFAGIGGIRRGFESIGGQC  251 ( 93)  132 ( 89)  mdcm
VPKNFNGVWVEPFMGTGVVAFNVAPKDALL  209 ( 39)  151 ( 31)  mecorv
RHLPKGECLVEPFVGAGSVFLNTDFSRYII  213 ( 35)  127 (  7)  mdam
LIPKTTNRYFEPFVGGGALFFDLAPKDAVI  213 ( 43)  144 ( 14)  mdpnii
KKLVFDDISVSSFCGDGDFRSSESIDLLKK  207 (116)  153 ( 64)  mecori
SSSKPNDVVLDPFFGTGTTGAVAKALGRNY  221 (220)  170 ( 55)  mhinfi
ASTKEGDYILDPFVGSGTTGVVAKRLGRRF  223 (215)  164 ( 40)  mdpna
IGTKKDSLVLDFFAGSGTTAEAVAYLNEKD  224 (436)  166 (163)  mecop1
FACDGEGIVLDFFAGSGTTAHTVFNLNNKN  224 (442)  166 (163)  mecop15
NVKDDAHVFMDLFSGTGIVGENFKKDYQVL  232 ( 33)  143 (  3)  mfoki
-----------------------------------------------------
VDTVGFDNKINYFHANGKPLDISLAKGLWV  176 (426)  160 ( 38)  mpsti
SVFECCKQTNRNFIGCDLIFGDDENEQD..  161 (213)  151 ( 23)  mhhaii
KMKYLTQDTTKFFTGKALLIKTASAGKRGG  168 (294)  156 (267)  mcvibiii
EVQWRGQGVINPFAESGGLVDLGEYPRLRR  186 (352)  166 ( 27)  mpaervii
FPNKKVSAVVIRFQKSGKGLSLWDTQESES  161 (208)  157 ( 67)  mtaqi
SHFPKYNRFVDLFCGGLSVSLNVNGPVLAN  191 ( 32)  129 ( 22)  mt4
MLTEPDDLVVDIFGGSNTTGLVAERESRKW  190 (273)  166 ( 97)  mpvuii

.....LF.G.G.....G... Most common amino acids
```

FIG. 4. Alignment of 30 DNA methylases about the motif F.G.G. See legend to Fig. 3 for explanation of the columns of numbers. The F.G.G pattern is repeated at position 143 in sequence M.*Ngo* II, but the alignment score for that position is less than second-best. In M.*Sin* I the F.G.G pattern is at residue position 343, but the best alignment is at residue 81. MOTIF parameter settings were [23,1,5] and the cpu time was 18 sec.

```
Sequence: HRG in 15 proteins (0 duplicates), Score 48

     +      +H +R*+ +  ++  *  + ++++ +G*+ +  +++*++  +
LERTGIELPAGQLTHVLRHTFASHFMMNGGNILVLQRVLGHTDIKMTMRYAHFAP  383 (277)  167 (187)  INT186
RSIWNGTGMAEWSLHDMRRTIATNLSELGCPPHVIEKLLGHQMVGVMAHYNLHDY  332 (352)  204 ( 89)  INT80
MKAIKPDLPMGQATHALRHSFATHFMINGGSIITLQRILGHTRIEQTMVYAHFAP  375 (269)  152 ( 59)  INTP2
RAALKRAGIEDFRFHDLRHTWASWLVQAGVPISVLQEMGGWESIEMVRRYAHLAP  351 (314)  152 (239)  INTP22
VRRIVKRTGIEFTSHMLRHTHATQLIREGWDVAFVQKRLGHAHVQTTLNTYVHLS  315 (302)  186 (195)  TNPA554
TSGGNAGLSLEIHPHMLRHSCGFALANMGIDTRLIQDYLGHRNIRHTVWYTASNA  361 (141)  170 (  7)  FIMB
DAGIEAGTVTQTHPHMLRHACGYELAERGADTRLIQDYLGHRNIRHTVRYTASNA  366 (136)  172 ( 38)  FIME
AKDDSGQRYLAWSGHSARVGAARDMARAGVSIPEIMQAGGWTNVNIVMNYIRNLD  283 (187)  167 (136)  CREP1
AAADGVTFSVPVTPHTFRHSYAMHMLYAGIPLKVLQSLMGHKSISSTEVYTKVFA  357 ( 60)  158 ( 15)  FPLORF
........MGRRRSHERRDLPPNLYIRNNGYYCYRDPRTGKEFGLGRDRRIAITE  211 (  7)  107 (  6)  INTLAM
VNRIFKSYSNVITPHQLRHFFCTNAIEKGFSIHEVANQAGHSNIHTTLLYTNPNQ  338 (234)  163 ( 71)  TN4430
TEKYTRAFDEKKSPHKLRHTYATNHYNENKDLVLLANQMGHNSMETTALYTNIDD  331 ( 17)  103 (  8)  BTHBT13
NIVDKSGEIYRFHAHAFRHTVGTRMINNGMPQHIVQKFLGHESPEMTSRYAHIFD  357 (465)  170 (126)  TNPB554
........MGRRRSHERRDLPPNLYIRNNGYYCYRDPRTGKEFGLGRDRRIAITE  211 (  7)  107 (  6)  HKINT
VLRAEIELPKGQLTHVLRHTFASHFMMNGGNILVLKEILGHSTIEMTMRYAHFAP  378 (280)  158 (187)  HPINT1
-------------------------------------------------------------------------------
SGLWSDDAIERQLSHSERNNVRAAYIHTSEHLDERRLMMQWWADYLDMNRNKYIS  170 (375)  169 (348)  INTP4
APYSIFAIKNGPKSHIGRHLMTSFLSMKGLTELTNVVGNWSDKRASAVARTTYTH  183 (305)  159 (258)  FLP
PASPIFAIKHGPKSHLGRHLMNSFLHKNELDSWANSLGNWSSSQNQRESGARLGY  165 (299)  149 (276)  FLPSB3

.............H.LRH..A......G......Q..LGH..I..T..Y.....  Most common amino acids
```

FIG. 5. A large sequence motif in the DNA integrases. See legend to Fig. 3 for explanation of the columns of numbers. MOTIF parameter settings were [15,0,24] and the cpu time was 214 sec.

appear. Shuffled sequences begin to produce a random motif background at a setting of [14,1,5].

**A Sequence Motif in the DNA Integrases.** Eighteen DNA integrases were examined. A single large motif spanning at least 36 residues is found in 15 of the proteins (Fig. 5). Partial alignments are found in the other 3 proteins. Four of the most highly conserved residues in the motif are histidine, arginine, glycine, and tyrosine, separated by distances d1 = 2, d2 = 21, and d3 = 9. Two of the sequences, intlam and hkint, are misaligned in the main block. The correct alignments of these two sequences, found by setting the significance parameter lower to pick up various submotifs of the larger motif, place the histidine residue at position 309 in each sequence. The misalignment is due to a single residue deletion that reduces d2 to 20. This separates one part of the pattern from the other. It just happens that an optimal fit to the first half of the pattern occurs in the amino-terminal portion of these two proteins. In addition, the sequence tpna554 has d3 = 10 but is still correctly aligned. Since no allowance is made for gaps in the current version of the program, sequences will sometimes be misaligned. The greatest utility of the program is in discovering motifs, not in correctly aligning every sequence. Some sequences will contain degenerate versions of the motif or gaps that will lower the alignment score sufficiently that incorrect patterns are chosen.

**Sequence Motifs in Reverse Transcriptases.** Thirty-three reverse transcriptases were analyzed. Three motifs were found at a setting of [21,2,5]. Y.DD (score 31) occurs in 28 of the proteins and the 50% degenerate motif is Y.DDIL (Fig. 6). The motif PQG (score 27) is found in 23 sequences and the 50% degenerate motif is P..RY.W.VLPQG.K.SP..FQ.....L. The motif R...D....N (score 35) is found in 20 proteins and the 50% degenerate motif is KK..GKWR.L.DLR..N..T.....PG. These motifs are found in the same relative order in all the sequences in which they occur.

## DISCUSSION

The discovery that eukaryotic genes are in pieces that are spliced together at the mRNA level (19) led to the proposal that the gene pieces (exons) correspond to functional domains and that new protein genes evolve by recombining functional pieces from other genes (20). If, in evolution, a particular exon were to be incorporated into a group of proteins, those proteins might over time conserve certain amino acids specific or crucial to a particular function carried by the exon, whereas other amino acids within the exon that were of lesser importance might undergo extensive mutation. As a consequence, any homologies would be obliterated except those retained at the few crucial positions. The conserved residues would constitute a sequence motif common to the group of proteins and correlated to the particular function. With these ideas in mind, we and others have sought a means to compare large, functionally related groups of proteins for highly conserved residue patterns.

As test cases for our program, we assembled several large groups of related proteins in which conserved amino acid motifs had been identified. The DNA methylases found in restriction and modification systems have been extensively compared (7–11). Our program readily finds the DPPY motif common to the adenine methylases and it finds eight of the cytosine methylase motifs reported by Posfai *et al.* (7). Two additional motifs reported by these authors are too degenerate for our program to find. By comparing all of the methylases, including one $N^4$-cytosine methylase, we find the motif F.G.G in 23 out of 30 sequences, and there are partial fits in most of the remaining sequences. The complete 50% degenerate motif is LF.G.G.....G. We speculate that the F.G.G motif might be part of the binding domain for *S*-adenosylmethionine, the common methyl donor for all these methylases. The F.G.G motif has also been reported by Lauster (21). In addition, we readily find the major motifs reported by others for the integrases (4) and the reverse transcriptases (6).

Global homologies are present in some of the large groups—for example, in the reverse transcriptases and the cytosine methylases. Our program rapidly locates and aligns segments of each sequence with respect to common motifs. The blocks of aligned segments can serve as anchor points for further alignment as has been pointed out by Posfai *et al.* (7). Optimal multisequence alignment should be possible in these interblock regions by using algorithms of the type developed by Lipman *et al.* (15) and Posfai *et al.* (7). Thus our approach in combination with these methods might lead to an efficient, relatively fast program for global alignment of >10 sequences.

```
Sequence: YDD in 28 proteins (0 duplicates), Score 31

      +  +++Y*DD*** +++  +
PIRQAFPQCTILQYMDDILLASPSHEDLLL  271 (161)  171 (122)  APOL
PMRKMFPTSTIVQYMDDILLASPTNEELQQ  272 (161)  171 (122)  TLPH
QVSAAFSQSLLVSYMDDILIASPTEEQRSQ  268 (161)  164 (105)  POBO
KVRHAWKQMYIIHYMDDILIAGKDGQQVLQ  272 (160)  175 ( 81)  SADP
TVRDKYQDSYIVHYMDDILLAHPSRSIVDE  266 (162)  189 (154)  RMTV
PIRKQFTSLIVIHYMDDILICHKELDVLQK  269 (162)  197 (142)  HART
PVREKFSDCYIIHYIDDILCAAETKDKLID  250 (160)  166 (156)  HERV
PLRLKHPSLCMLHYMDDLLLAASSHDGLEA  263 (160)  145 ( 48)  RVPO
PFKKQNPDIVIYQYMDDLYVGSDLEIGQHR  249 (163)  184 (124)  RTAV
PFRKANKDVIIIQYMDDILIASDRTDLEHD  270 (161)  179 (122)  HIV2
PFRERYPEVQLYQYMDDLFVGSNGSKKQHK  260 (161)  179 ( 35)  EAVP
GWIEEHPMIQFGIYMDDIYIGSDLGLEEHR  243 (162)  172 (123)  VLVP
DFRIQHPDLILLQYVDDLLLAATSELDCQQ  261 (162)  148 ( 86)  LVP1
KFPTRDLGCVLLQYVDDLLLGHPTAVGWPR  252 (162)  168 ( 73)  ERVH
VVRRAFPHCLAFSYMDDVVLGAKSVQHLES  252 (161)  196 ( 99)  HEPB
DILRPLLNKHCLVYLDDIIVFSTSLDEHLQ  255 (155)  176 ( 10)  176D
NILRPLLNKHCLVYLDDIIIFSTSLTEHLN  255 (156)  175 ( 89)  297D
DVLREQIGKICYVYVDDVIIFSENESDHVR  249 (157)  169 (104)  GYPD
IAFSGIEPSQAFLYMDDLIVIGCSEKHMLK  248 (157)  193 ( 35)  412D
MADTFRDLRFVNVYLDDILIFSESPEEHWK  259 (145)  192 (131)  TYS2
DEAFRVFRKFCCVYVDDILVFSNNEEDHLL  252 (153)  170 (119)  CAMV
NSHSNQYSKYCCVYVDDILVFSNTGRKEHY  251 (154)  170 (119)  CERV
LRMLRDINVSVIAYLDDLLIVGSTKEECLS  261 (159)  182 ( 67)  DIRS
SNIILHKEIKFNAYADDFFLIINFNKNTNT  224 (160)  203 ( 53)  IFAC
IHGIQSNKLMYVRYADDWIVAVNGSYTQTK  222 (168)  184 ( 92)  C2IS
YNDPNFKRMKYVRYADDILIGVLGSKNDCK  246 (164)  187 ( 65)  QXB1
VKELFKRYDELIMYADDGILCRQDPSTPDF  218 (162)  217 ( 50)  MAUP
DKGNINENIYVLLYVDDVVIATGDMTRMNN  247 (157)  169 ( 41)  COPD
-------------------------------------------------------
EVKLSLFADDMIVYLENPIVSAQNLLKLIS  221 (180)  194 (173)  HLIN
LKSGTRQGCPLSPYLFNIVLEVLARAIRQQ  204 (139)  183 ( 58)  MLIN
SRSPIQATAQLALYLIDIKKWLSDWRIKVN  185 (183)  185 ( 68)  FE2D
QRLAEVPLLQHGFFADDLTLLARHTERDVI  177 (160)  163 ( 62)  INGT
SCVFKNSQVTICLFVDDMVLFSKNLNSNKR  207 (162)  192 ( 78)  TYEY1

.............Y.DDIL.....  Most common amino acids
```

FIG. 6. The most significant sequence motif among 33 reverse transcriptases is Y.DD. See legend to Fig. 3 for explanation of the columns of numbers. The first unmatched sequence (below dotted line) has a second-best alignment of FADD at position 173; the second and third unmatched sequences have LADD (173) and FADD (160), but they score less than second-best.

A shortcoming of our program is its inability to deal with gaps. Clearly not all residue spacings in a motif will be rigidly conserved. The integrases provide an example in which at least three of the sequences have single-residue deletions or additions within the motif. One sequence is still correctly aligned, the other two are not. However, alignment is correct in several of the submotifs of the larger motif, and by inspection of these we are able to find the correct alignments. Our program will successfully find motifs if the number of sequences is large and if only a few sequences have gaps located in the motif. MOTIF will fail if gapping is extensive or if the motif is very degenerate. We currently do not see a way to allow for gaps without significantly increasing background and computation time.

In designing MOTIF, we found that 2-amino acid patterns gave too high a random background of matches and 3-amino acid patterns gave a satisfactory background, whereas 4-amino acid patterns, while decreasing the background, put too much constraint on the type of motif found. Although the background of random motifs is fairly significant with 3-amino acid patterns (see Fig. 2 and *Results*), we have satisfactorily dealt with this in two ways. First, we have incorporated a parameter that sets a limit on the number of allowed internal repetitions of the pattern. Real motifs tend to occur singly in each sequence, whereas background motifs are dispersed randomly over the total sequence length; consequently they are normally present in more than one copy in some sequences while being absent from others. However, in using MOTIF it is not wise to always set the repeat parameter to zero. Some proteins have internal duplications (e.g., M.*Fok* I in the adenine methylase group), and there is also chance of occasional random internal repetitions of a real motif pattern. Second, we have incorporated a sequence-shuffling procedure to permit the random background of spurious motifs to be empirically determined for any group of sequences. Thus when a motif is found, it can be presumed to be real if no spurious motifs are found after shuffling using the same parameter settings.

Spurious motifs, which generally occur at low significance parameter settings, can in most cases be identified by their low total scores and by their amino acid content. The low score derives from the minimal level of surrounding homology. Amino acid content is also a useful discriminator. Real motifs often contain the less common amino acids, while spurious motifs contain the most common ones. By applying the shuffling routine to a few sequence groups, it quickly becomes evident that background motifs consist largely of the most common amino acids. MOTIF displays the frequencies of each of the amino acids in the protein group to aid in identifying the common amino acids. With these characteristics in mind, it is generally possible to differentiate between real and spurious motifs. We believe that MOTIF will prove most useful in discovering motifs in large groups of poorly conserved sequences that are not amenable to analysis by current multisequence alignment procedures. The program is available on request.

This work was supported by American Cancer Society Grant NP 52282. H.O.S. is an American Cancer Society Research Professor. S.C. has a Junior Faculty Research Award from the American Cancer Society and is supported by National Institutes of Health Grant GM42140.

1. Fry, D. C., Kuby, S. A. & Mildvan, A. S. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 907–911.
2. Hunter, T. (1987) *Cell* **50**, 823–829.
3. Bashford, D., Chothia, C. & Lesk, A. M. (1987) *J. Mol. Biol.* **196**, 199–215.
4. Argos, P., Landy, A., Abremski, K., Egan, J. B., Haggard-Ljungquist, E., Hoess, R. H., Kahn, M. L., Kalionis, B., Narayana, S. V. L., Pierson, L. S., III, Sternberg, N. & Leong, J. M. (1986) *EMBO J.* **5**, 433–440.
5. Romisch, K., Webb, J., Herz, J., Prehn, S., Frank, R., Vingron, M. & Dobberstein, B. (1989) *Nature (London)* **340**, 478–482.
6. Doolittle, R. F., Feng, D.-F., Johnson, M. S. & McClure, M. A. (1989) *Q. Rev. Biol.* **64**, 1–30.
7. Posfai, J., Bahwat, A. S., Posfai, G. & Roberts, R. J. (1989) *Nucleic Acids Res.* **17**, 2421–2435.
8. Chandrasegaran, S. & Smith, H. O. (1988) in *Structure and Expression*, eds. Sarma, R. H. & Sarma, M. H. (Adenine, Guilderland, NY), Vol. 1, pp. 149–156.
9. Sznyter, L. A., Slatko, B., Moran, L., O'Donnell, K. H. & Brooks, J. E. (1987) *Nucleic Acids Res.* **15**, 8249–8266.
10. Caserta, M., Zacharias, W., Nwankwo, D., Wilson, G. G. & Wells, R. D. (1987) *J. Biol. Chem.* **262**, 4770–4777.
11. Som, S., Bhagwat, A. S. & Friedman, S. (1987) *Nucleic Acids Res.* **15**, 313–332.
12. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227**, 1435–1441.
13. Argos, P. (1985) *EMBO J.* **4**, 1351–1355.
14. Zalkin, H., Argos, P., Narayana, S. V. L., Tiedeman, A. A. & Smith, J. M. (1985) *J. Biol. Chem.* **260**, 3350–3354.
15. Lipman, D. J., Altschul, S. F. & Kececioglu, J. D. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 4412–4415.
16. Murata, M., Richardson, J. S. & Sussman, J. L. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 3073–3077.
17. Dayhoff, M. O., Eck, R. V. & Park, C. M. (1972) *A Model of Evolutionary Change in Proteins* (Natl. Biomed. Res. Found., Washington, DC), Vol. 5, p. 89.
18. Tao, T., Walter, J., Brennan, K. J., Cotterman, M. M. & Blumenthal, R. M. (1989) *Nucleic Acids Res.* **17**, 4161–4175.
19. Berget, S. M., Moore, C. & Sharp, P. A. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 3171–3175.
20. Gilbert, W. (1978) *Nature (London)* **271**, 501.
21. Lauster, R. (1989) *J. Mol. Biol.* **206**, 313–321.