Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 12091–12095, December 1994
Evolution

Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks

ROMAN L. TATUSOV, STEPHEN F. ALTSCHUL, AND EUGENE V. KOONIN*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894

Communicated by Charles R. Cantor, August 9, 1994

ABSTRACT **We describe an approach to analyzing protein sequence databases that, starting from a single uncharacterized sequence or group of related sequences, generates blocks of conserved segments. The procedure involves iterative database scans with an evolving position-dependent weight matrix constructed from a coevolving set of aligned conserved segments. For each iteration, the expected distribution of matrix scores under a random model is used to set a cutoff score for the inclusion of a segment in the next iteration. This cutoff may be calculated to allow the chance inclusion of either a fixed number or a fixed proportion of false positive segments. With sufficiently high cutoff scores, the procedure converged for all alignment blocks studied, with varying numbers of iterations required. Different methods for calculating weight matrices from alignment blocks were compared. The most effective of those tested was a logarithm-of-odds, Bayesian-based approach that used prior residue probabilities calculated from a mixture of Dirichlet distributions. The procedure described was used to detect novel conserved motifs of potential biological importance.**

With the rapid growth of genome sequence information and the inability of current experimental techniques to generate functional information at an equal pace, computer-assisted analysis is becoming a focus of modern biology. While the largest complete genome sequences now available are those of organelles and large DNA viruses, within a few years we will be able to analyze the complete genomes of such organisms as *Escherichia coli* and yeast (1). The extent to which these sequences prove valuable will depend heavily upon the power of computer analysis.

The study of a new sequence generally begins with a database similarity search (reviewed in ref. 2), such as that performed by the FASTA (3) or BLAST (4–6) algorithms. Once a group of similar sequences has been found, local multiple alignment methods may be used to extract common patterns (e.g., refs. 7–10). These methods may uncover weak but potentially functionally important conservation that was undetectable by simple database search.

The database may be scanned for sequence motifs that have been extracted from multiple alignments and are generally associated with a particular function. Motifs may be represented as specific patterns of required or permitted amino acids (e.g., refs. 11–13). An alternative representation is provided by position-dependent weight matrices or profiles (e.g., refs. 8, 10, and 14–21). These may be generated either from gapped multiple alignments (17) or from aligned blocks of relatively short protein segments (typically between 12 and 35 residues) containing no gaps (10, 20). Local multiple alignments reveal the boundaries of conserved sequence regions and the relative importance of various residues within them. By taking advantage of this information, motif search methods are capable of detecting subtle sequence similarities. They appear to be the tools of choice for identifying members of protein families and for classifying protein sequences (e.g., ref. 22).

We describe here a statistically based approach to the identification within a sequence database of protein segments related to an ungapped alignment block. We then apply this method in a semiautomatic strategy for delineating protein families, starting from a single sequence or a group of related sequences.

From Alignment Blocks to Weight Matrices

The simplest weight matrices for database searching, and those we will study here, are ones that do not allow gaps and that must be matched across their entire lengths (8, 10, 15, 18, 20). Such matrices are generated from alignment blocks which consist of N ungapped sequence segments of length L. A protein weight matrix, W_{jk}, generated from such a block will have 20 rows (one for each possible residue) and L columns. Given an alignment block, there are many possible ways in which the scores of the corresponding weight matrix may be calculated. We have investigated four methods.

Method A: Average-Score Method. For pairwise protein sequence comparison, many different sets of amino acid substitution scores, such as the PAM (23–25) or BLOSUM (26, 27) matrices, have been proposed. Let the scores in such a substitution matrix be S_{ij}, and suppose that amino acid i occurs C_{ik} times within column k of the given alignment block. Perhaps the simplest way to generate a position-dependent weight matrix from the scores S_{ij} is to average them: let $W_{jk} = (\Sigma_{i=1}^{20} C_{ik} S_{ij})/N$. This method is essentially that proposed by Gribskov *et al.* (17) for the calculation of profiles.

Method B: Bayesian Prediction Method. Theory supports specifying weight matrix scores to be $W_{jk} = \log(q_{jk}/p_j)$, where q_{jk} is the probability for residue j to occur in motif position k, and p_j is the "background" probability of residue j (8, 10, 15, 16, 19, 21). As the number of essentially independent segments grows, estimates of q_{jk} should converge to C_{jk}/N, but this is a poor formula for small N (10, 15, 16). The simplest Bayesian prediction approach estimates q_{jk} as $(C_{jk} + Bp_j)/(N + B)$ (10). The parameter B may be thought of as "pseudocounts," allotted among the residues in proportion to the p_j; empirically, choosing $B \approx \sqrt{N}$ has proven efficacious (10).

Method C: Data-Dependent Pseudocount Method. Method B unfortunately ignores amino acid interrelationships. A simple yet *ad hoc* way around this problem is for the residue pseudocounts to depend, via a substitution matrix S_{ij}, upon the observed data. Formally, the number of pseudocounts, Bp_j, for residue j in motif position k may be multiplied by $(\Sigma_{i=1}^{20} C_{ik} e^{\lambda S_{ij}})/N$, where λ is the natural scale for matrix S (28). Pseudocounts for residues similar to those observed are augmented at the cost of residues dissimilar. This method may

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.

*To whom reprint requests should be addressed.

approximate using the matrix S_{ij} when $N = 1$, and estimating $q_{jk} = C_{jk}/N$ for large N.

Method D: Dirichlet Mixture Method. The Bayesian prediction method above can be generalized by letting the prior probabilities for the q_{jk} be a mixture of multiple Dirichlet distributions, as opposed to just one (21). This is a mathematically rigorous way to graft the notion of amino acid relationships onto method B. In making the pseudocounts for each residue data-dependent, the approach is similar to the *ad hoc* method C.

Given a weight matrix produced by any of these methods, a protein database is searched by sliding the matrix along each sequence and computing a score for every segment of length L by summing the appropriate matrix elements.

Statistics of Weight Matrix Database Searches

It is possible to calculate precisely the distribution of segment scores implied by the type of weight matrix described above. We assume that the matrix elements are all integers—scaling and rounding can produce any desired level of precision. The score for a segment of length L may then assume only integral values. Given a protein model in which every amino acid is chosen independently with background residue probabilities p_j, the probability for every possible segment score may then be calculated straightforwardly (14, 29). In brief, assume that the probability distribution $P_{a-1}(x)$ for a matrix consisting of the first $a - 1$ columns of W is known. Then, inductively, $P_a(x) = \Sigma_{j=1}^{20} P_{a-1}(x - W_{ja})p_j$. The probabilities for the possible scores from the first column of W may of course be calculated directly from the p_j.

Although the scores of overlapping segments are not independent, the theoretical number of database segments expected to achieve a score x is simply $P_L(x)$ times the number of segments examined. It is possible to generate an empirical score distribution by scanning the database with the matrix and collecting all the resulting segment scores. Furthermore, a theoretical probability distribution for scores from true positive protein segments can be computed by using the residue frequencies q_{jk} in place of the p_j in the calculation described above.

A comparison is shown in Fig. 1 of the theoretical and empirical score distributions for a matrix derived from a block conserved in a well-characterized superfamily of RNA-dependent RNA polymerases. First, the theoretical distribution differs significantly from a normal distribution with the same mean and standard deviation (curves 1 and 3 in Fig. 1*A*). Specifically, the normal distribution underestimates the probability of high scores and thus provides a poor approximation for weight matrix database searching. Second, many more high scores are found empirically than predicted by the right-hand tail of the theoretical distribution (curves 1 and 2 in Fig. 1*A* and curves indicated in Fig. 1*B*). These scores are, however, reasonably well modeled by the theoretical curve for true positive segments (curve 4 in Fig. 1*A*). Anomalous high scores vanish when the columns of the matrix are shuffled (Fig. 1*B*).

The number of segments predicted to attain score above a given cutoff value gives the expected number, F, of false positives—i.e., segments that are similar to the block by chance. Subtracting F from the number of scores over the cutoff actually observed gives a predicted number, T, of true positives—i.e., segments that are similar to the block for biological reasons. The ratio $R = F/T$ gives the estimated odds that a segment with score over the cutoff is a false positive. Cutoff scores may alternatively be characterized solely by F. Odds ratios have the advantage that the same cutoff is appropriate for matrices representing both large and small protein families; we routinely use R to set cutoffs in our search procedure.

We have developed a program called MoST (Motif Search Tool) that takes as input an alignment block, constructs a corresponding weight matrix, scans a protein database with this matrix, produces theoretical and empirical score distributions, retrieves segments exceeding an appropriate cutoff score, and assesses their statistical significance. The program incorporates as well two important semiheuristic procedures. First, in the matrix construction stage, it employs a simple weighting procedure that groups segments with reference to their percent identity. Second, to remove compositionally biased segments that frequently yield spurious hits in database searches, the segments retrieved are filtered with the SEG program (2, 47).

The execution time for MoST is proportional to the length of the database and the length L of the input alignment block but is essentially independent of the number of sequences in the block. Using a single processor SGI workstation to scan the Swiss-Prot database (30), Release 26 (10,875,091 residues), required ≈1 sec per weight matrix column.

Relative Discriminating Power of Weight Matrices

To measure how well a weight matrix distinguishes a biological motif from background noise, one needs a reliable *a priori* division of the database into segments that do and do not instantiate the motif. For this purpose we used several

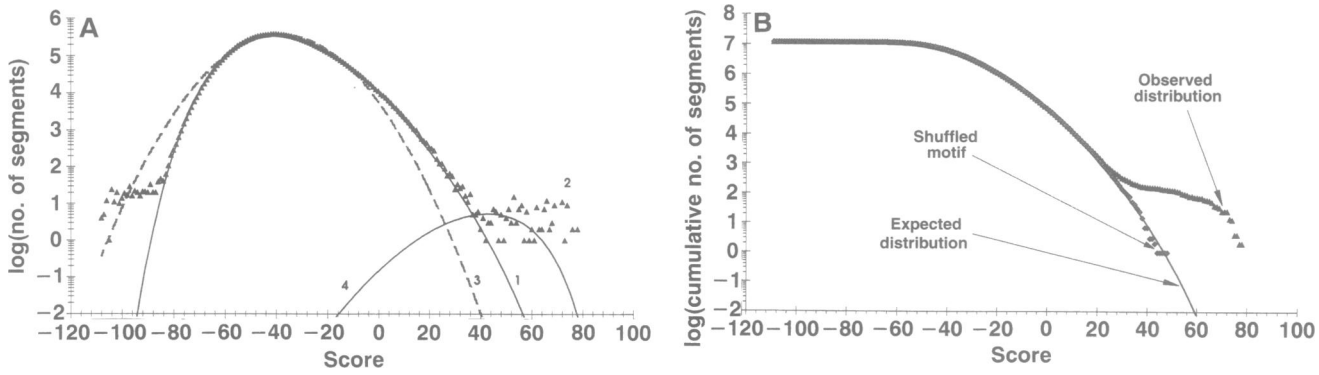

FIG. 1. Theoretical and empirical score distributions for segments from Swiss-Prot (30), Release 27. The *y*-axis scales are logarithmic. (*A*) Score distributions for a weight matrix derived by the Dirichlet mixture method from the positive-strand and double-stranded RNA virus RNA-dependent RNA polymerase (RdRp) motif V (31). The alignment block consists of 13 segments of 18 residues each, representing different groups of polymerases. All RdRps contain a set of four conserved motifs that unequivocally define the superfamily (31). There are 175 members of the RdRp superfamily in Swiss-Prot, Release 27. Curves: 1, theoretical distribution assuming background residue probabilities p_j; 2, empirical distribution; 3, normal distribution with same mean and standard deviation as curve 1; 4, theoretical distribution assuming position-dependent residue frequencies q_{jk} and 175 instances of the motif. (*B*) Cumulative distributions (number of scores $\geq x$) for the weight matrix from RdRp motif V.

Table 1. Discrimination power (no. of false positives, E) of position-dependent weight matrices calculated by various methods

	Helicases I		Helicases II		RdRp		DdDp		UvrA-related
Method	Motif V (n = 85)	Motif VI (n = 84)	Motif V (n = 128)	Motif VI (n = 125)	Motif IV (n = 175)	Motif V (n = 175)	Motif A (n = 51)	Motif B (n = 51)	ATPases, motif II (n = 147)
A	39	42	45	20	33	27	5	8	21
B	40	39	32	28	28	27	7	8	27
C	34	36	30	28	27	27	6	7	26
D	29	27	30	18	22	23	3	8	21
Control	49	48	41	30	34	40	9	3	26

For each alignment block and method for calculating segment scores, a cutoff was chosen for which the number of false positives, E, equaled the number of false negatives. The table gives E for each block and method; lower E corresponds to better discrimination. The alignment blocks for motifs V and VI from the two distantly related superfamilies of (putative) helicases (reviewed in ref. 32) included, respectively, 9 and 10 segments from experimentally characterized helicases and nucleic acid-dependent ATPases. For helicase motifs V and VI, a different size of superfamily (n) is indicated because only putative helicase fragments containing one of the motifs are available in Swiss-Prot. The RNA-dependent RNA polymerase (RdRp) motifs were represented by blocks of 14 segments from different groups of positive-strand RNA virus and double-stranded RNA virus polymerases (31). The A and B motifs from the superfamily of DNA-dependent DNA polymerases (DdDp) (33) were represented by blocks of 10 segments from cellular and viral DdDp. Motif II from the superfamily of UvrA-related ATPases (34) included 11 segments from experimentally characterized ATPases.

protein families that have been studied in sufficient detail that a canonical list of true family members could be produced (Table 1).

When all database segments have been assigned scores by a weight matrix representing a given motif, any particular cutoff score will yield a certain number of false negatives and false positives. These numbers reflect the inevitable trade-off between sensitivity and selectivity. Given a correct classification of all segments, a convenient measure of the power of a matrix may be constructed by finding the cutoff at which the number of false positives, E, equals the number of false negatives (W. Pearson, personal communication). Clearly, the lower is E the greater is the discriminating power of the matrix, with E ideally equal to zero. These values could be determined only for protein superfamilies with precisely defined membership. Five such superfamilies, defined by a combination of functional information and comparative sequence analysis, were selected.

Using the MoST program and the measure E, we explored the relative discriminating power of weight matrices constructed by methods A–D described above. For each alignment block, we also evaluated a control procedure that assigns to every database segment its maximum pairwise score, using a standard amino acid substitution matrix, with any segment from the alignment block. This control helps determine to what extent weight matrices abstract useful information from alignments, information not available collectively in the constituent segments.

For a variety of test motifs, we scanned the Swiss-Prot database (30) by using the methods described above. The results in Table 1 show that weight matrices, particularly those constructed by method D, significantly outperform the control in most cases. No difference was observed only with very selective motifs—e.g., those for DNA polymerases (Table 1)—that showed high discriminating power with all methods. The information implicit in an alignment block thus appears to be better captured by weight matrices than by the raw collection of segments that comprise the block. This is particularly true for weight matrices constructed (as in method D) with reference to both observed amino acid counts and *a priori* knowledge of residue relationships.

Iterative Weight Matrix Search

When a database is scanned, the retrieval of any segments that were not in the original alignment block potentially brings useful information. An obvious way to exploit this information is to generate a new alignment block and associated weight matrix and to repeat the search. Iterative approaches have been applied to motif searches in several studies (19, 35–37), but we are unaware of any systematic investigation of their behavior.

We designed an iterative procedure, based on the MoST algorithm, that searches databases for related sequences. During each stage a weight matrix is generated from the current alignment block, the theoretical score distribution for the matrix is calculated, the database is scanned to produce an empirical score distribution, a cutoff score is chosen based upon either a fixed number F or odds ratio R of false positives, and segments with score exceeding this cutoff are added to the alignment block. The process is repeated until no new segments are found.

Fig. 2 shows the dynamics of segment retrieval in iterative searches seeded with the same block whose score distribution is shown in Fig. 1. For this and all other initial alignment blocks studied, the process always converged for sufficiently stringent cutoff scores ($R < 0.02$ for the great majority of cases). The number of iterations required varied by case. When motifs with known positive and negative sets were studied, the observed number of false positives generally agreed with the number predicted by theory, although later iterations sometimes yielded a greater proportion of false positives (Fig. 2).

To assess the performance in an iterative search of methods A–D for weight matrix construction, we repeated until convergence the above steps for each family in our test set (see Table 1). For most of the motifs, methods B–D retrieved

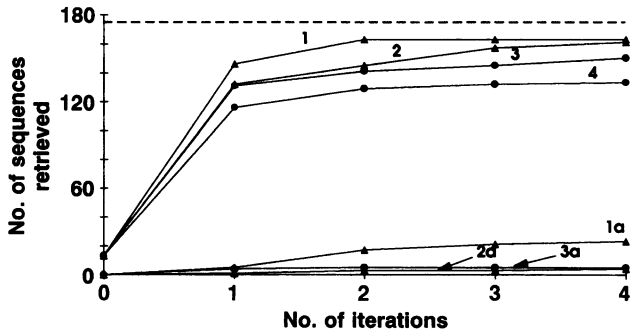

FIG. 2. Accumulation of segments related to an alignment block in iterative searches of Swiss-Prot. The search was done with the RNA-dependent RNA polymerase motif V. The broken line at 175 indicates the number of superfamily members. Dirichlet mixture method D for weight matrix generation: 1 and 1a, true and false positives for $R = 0.03$; 2 and 2a, true and false positives for $R = 0.01$. Average-score method A for weight matrix generation: 3 and 3a, true and false positives for $R = 0.03$; 4, true positives for $R = 0.01$ (there were no false positives).

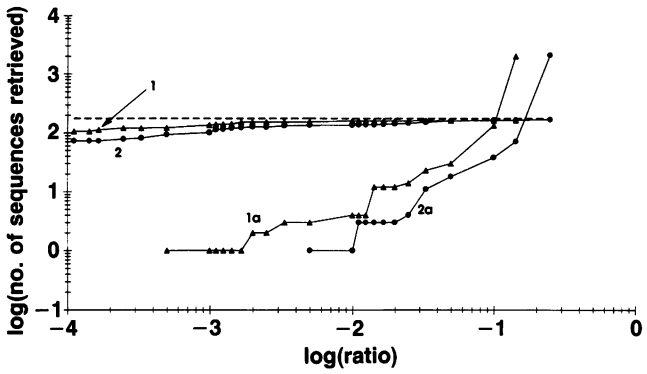

FIG. 3. The convergence of iterative search as a function of the cutoff ratio R. Weight matrices constructed from the RdRp motif V block were compared with Swiss-Prot. Both scales are logarithmic. Curves: 1 and 1a, true and false positives for the Dirichlet mixture method; 2 and 2a, true and false positives for the average-score method.

a considerably higher fraction of the known family members than either method A or the control (data not shown).

We also studied the behavior of iterative search as a function of the cutoff odds ratio R. At certain critical values, the number of false positives retrieved increased dramatically, changing the outcome of the iterative search process from convergence to divergence (Fig. 3). Convergence typically was not observed when an analogous iterative procedure was implemented for direct similarity search using the BLAST algorithm. Except for very high cutoffs, the number of the retrieved sequences grew explosively, with a majority of false positives even at early iterations (data not shown).

Our iterative procedure is flexible in the choice of cutoff values for successive iterations as well as in the method for weight matrix construction. We found that using a relatively high R(0.05–0.1) for the first iteration, and more stringent values subsequently, frequently led to improved results (data not shown). This was the case particularly when the original alignment block contained few sequences.

From Pairwise Database Searches to Alignment Blocks

The procedure described above uses an alignment block to seek a motif present in sequences throughout the database. Ideally, one would start with a single, uncharacterized sequence and construct a conserved block representative of the protein family to which the sequence belongs. In many cases, when only weak, perhaps statistically nonsignificant similarities are generated by a BLAST search, it is unclear which if any of the alignments is functionally relevant. One indication of possible relevance is the appearance of the same query sequence segment in different alignments (5), and this is one possible approach to the delineation of conserved blocks. We have developed a program called CAP (Consistent Alignment Parser), which constructs alignment blocks from BLAST search output. Alternatively, sequences with BLAST hits may be extracted from the database and analyzed by using a program for local multiple alignment such as MACAW (9) or the Gibbs sampler (10). The alignment blocks found by any of these methods can be used as input for a MoST search.

New Findings with Old Motifs and New Motifs from Uncharacterized Sequences

We analyzed various biologically important protein sequences by using the search strategy described above. Combined, iterative use of BLAST and MoST permitted the description of several new protein superfamilies, as well as a number of functional predictions (38–40). Below we describe two conserved domains found in proteins involved in DNA recombination, repair, and replication.

A highly conserved motif was found in bacterial RecR proteins and class I DNA topoisomerases. A BLAST database search with *E. coli* RecR protein (41, 42) found significant similarity only to the same protein in other bacteria. However, a subsequent search with MoST revealed a 25-residue motif shared by RecR proteins and class I topoisomerases. The same set of 12 sequences was retrieved from the nonredundant protein sequence database (NR, supported by the National Center for Biotechnology Information) by using R values as low as 0.01. A reciprocal test, beginning with only the topoisomerase segments, produced the same result. When each sequence of the set was compared with NR by BLAST, no significant new sequence similarities were found. Further analysis using MACAW showed that RecR proteins and topoisomerases shared a second, less obvious region of similarity (Fig. 4). The motif revealed by MoST has been previously described as one of the highly conserved regions in class I topoisomerases (43). An obvious common activity of RecR and topoisomerases is DNA binding, but the actual function of their shared sequence motifs remains to be established.

Another motif is shared by the *E. coli* repair protein RecJ, a subunit of DNA polymerase III (DnaE), and an uncharacterized putative protein encoded by open reading frame 30

```
RecR   Ec    82  ICVVESPADIYAIEQTGQF  29  RLAEEKITEVILATNPTVEGEATANYIAELC  P12727
RecM   Bs    81  ICVVQDPKDVIAMEKMKEY  29  RLQDDQVTEVILATNPNIEGEATAMYISRLL  P24277
RecR   Ml    81  VCVVEEPKDVQAVERTREF  29  RVDDVGITEVIIATDPNTEGEATATYLVRMV  L01263g

TrsI   Sa     3  LILCEKPSQAMDLSTVFAK  73  IFKENKIDEVIIATDPAREGENIAYKILNQL  L11998g
TOP1   Ec     4  LVIVESPAKAKTINKYLGS  73  KQLAEKADHIYLATDLDREGEAIAWRLREVI  P06612p
ORF1   Ef     3  VILAEKPSQALAYASALKQ  71  AELLKQANTIIVATDSDREGENIAWSIIHKA  PQ0259p
TOP1   Ssp    3  LVIVESPTKARTIRNYLPK  58  KDALKDADELILATDEDREGKVISWHLLQLL  S32158p
TOP1   Bf    68  --------------------  0  TIFDKRVKTIILATDAAAEGEYIGRNILYRL  S23866p
TOP3   Sc     3  LCVAEKNSIAKAVSQILGG  83  KREARNADYLMIWTDCDREGEYIGWEIWQEA  P13099
TOP3   Ec     2  LFIAEKPSLARAIADVLPK  67  KRFLHEASEIVHAGDPDREGQLLVDEVLDYL  P14294
RGYR   Sac  626  LLVVESPNKAKTISSFFSR  98  RNLAVEADEVLIGTDPDTEGEKIAWDLYLAL  L10651g
CONSENSUS        UxUvExpxxAxx&xxxxxx      xxxxxxxxxU&uatDxxxEGexxxxxUxxxu
                   aQ   U                          g N    q        a
```

FIG. 4. A conserved domain in RecR and class I topoisomerases (TOP). The alignment was generated with the MACAW program (9). The conserved block detected by an iterative MOST search of the protein NRDB is overlined. The consensus line indicates conserved residues. Uppercase letters indicate that all residues in the column are chemically similar; lowercase indicates that most are. U designates a bulky aliphatic residue (I, L, V, or M), and & designates a bulky hydrophobic residue (I, L, V, M, F, Y, or W). For each sequence, the number of N-terminal residues is indicated, as well as the number between the conserved blocks. TrsI is a transport protein from a *Staphylococcus aureus* (Sa) plasmid; RGYR is reverse gyrase from the archaeon *Sulfolobus acidocaldarius* (Sac). Other abbreviations: Ec, *E. coli*; Bs, *Bacillus subtilis*; Ml, *Micrococcus luteus*; Ef, *Enterococcus faecalis*; Ssp, *Synechococcus* sp.; Bf, *Bacillus firmus* (incomplete topoisomerase I sequence); Sc, *Saccharomyces cerevisiae*. The Swiss-Prot, Protein Identification Resource (PIR) (p), or GenBank (g) accession number is given for each sequence.

```
RecJ  Ec   90 LSVLAMRSLGCSNIDYLVPNRFEDGYG 26 HAGVEHARSLGIPVIVTDHHLPGD 401 P21893
DnaE  Ec   83 LTVLAANNTGYQNLTLLISKAYQRGYG 73 HAAVELAEARGLPVVATNDVRFID 954 P10443
ORF30 P2   19 FAVLAFFSFGKSNLMRLIAHYFNFGYS 85 EQSVIVRDATGIPYKNMMYYVYSD 106 U02597g
consensus     &.VLA....G..NU..LU...@..GYG    ...V....A.GUP&.........D
                                         S          S
```

FIG. 5. Putative exonuclease motif. The overlined proximal motif was identified by using the combination of BLAST and MOST as described in the text. The additional, distal motif was detected with MACAW (9). The consensus shows amino acid residues that are conserved in all the three aligned sequences. For other details and designations, see legend to Fig. 4.

(ORF30) from bacteriophage P2 (Fig. 5). A BLAST search with the RecJ sequence revealed only very limited, not statistically significant similarity with DnaE. When a 27-amino acid region that was conserved between RecJ and DnaE was used for a MOST search, a related segment was identified in the P2 protein. Each pair of sequences from this motif identified the third sequence with R value below 0.01, whereas no other sequences were selected from the database even with R = 0.1. These observations show that, at least in some cases, an alignment of only two sequences may produce a sensitive position-dependent weight matrix that can be used to identify a specific, conserved motif. RecJ is a 5′–3′ single-stranded-DNA exonuclease (44, 45). Therefore we predict that the N-terminal domains of DnaE and the P2 ORF30 product, which is not essential for bacteriophage propagation in *E. coli* (46), possess a similar exonuclease activity, and the conserved motif may be a part of the active center.

Conclusions

We have described an approach to protein database analysis using weight matrices derived from alignment blocks. The statistical distribution of matrix scores provides a measure for assessing database segments. For each motif explored, iterative database searches converged on an aligned block of segments containing the motif. We used this approach to evaluate the discriminating power of different methods for computing weight matrices. The use of Dirichlet mixture priors (21) generally was the most effective. The initial blocks used to seed this approach can be generated from the output of a standard database similarity search program, such as BLAST, by parsing consistent segments from the alignments reported. Thus, in principle, our strategy allows the construction from a single uncharacterized sequence of conserved motifs characteristic of a protein superfamily. Using this approach, we identified several motifs of potential functional importance that were not detectable by direct database search. Source code and executable versions of the programs MOST and CAP are available from the authors upon request.

We thank Dr. David Lipman for constant encouragement and helpful discussions; Drs. Warren Gish, David Landsman, David Lipman, and John Wootton for critical reading of the manuscript; Dr. David Haussler for his program for estimating residue probabilities based upon Dirichlet mixtures; and Dr. William Pearson for providing unpublished results.

1. Bork, P., Ouzounis, C. & Sander, C. (1994) *Curr. Opin. Struct. Biol.* **4**, 393–403.
2. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6**, 119–129.
3. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.
4. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403–410.
5. Altschul, S. F. & Lipman, D. J. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 5509–5513.
6. Gish, W. & States, D. J. (1993) *Nat. Genet.* **3**, 266–272.
7. Posfai, J., Bhagwat, A. S., Posfai, G. & Roberts, R. J. (1989) *Nucleic Acids Res.* **17**, 2421–2435.
8. Stormo, G. D. & Hartzell, G. W., III (1989) *Proc. Natl. Acad. Sci. USA* **86**, 1183–1187.
9. Schuler, G. D., Altschul, S. F. & Lipman, D. J. (1991) *Proteins* **9**, 180–190.
10. Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C. (1993) *Science* **262**, 208–214.
11. Hodgman, T. C. (1989) *Comput. Appl. Biosci.* **5**, 1–13.
12. Smith, R. F. & Smith, T. F. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 118–122.
13. Bairoch, A. (1993) *Nucleic Acids Res.* **21**, 3097–3103.
14. McLachlan, A. (1983) *J. Mol. Biol.* **169**, 15–30.
15. Schneider, T. S., Stormo, G. D., Gold, L. & Ehrenfeucht, A. (1986) *J. Mol. Biol.* **188**, 415–431.
16. Berg, O. G. & von Hippel, P. H. (1987) *J. Mol. Biol.* **193**, 723–750.
17. Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 4355–4358.
18. Stormo, G. D. (1988) *Annu. Rev. Biophys. Biophys. Chem.* **17**, 241–263.
19. Dodd, I. & Egan, J. B. (1990) *Nucleic Acids Res.* **18**, 5019–5026.
20. Henikoff, S. & Henikoff, J. G. (1991) *Nucleic Acids Res.* **19**, 6565–6572.
21. Brown, M., Hughey, R., Krogh, A., Mian, I. S., Sjölander, K. & Haussler, D. (1993) in *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, eds. Hunter, L., Searls, D. & Shavlik, J. (AAAI, Menlo Park, CA), pp. 47–55.
22. Koonin, E. V., Bork, P. & Sander, C. (1994) *EMBO J.* **13**, 493–503.
23. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (Natl. Biomed. Res. Found., Washington, DC), Vol. 5, Suppl. 3, pp. 345–352.
24. Schwartz, R. M. & Dayhoff, M. O. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (Natl. Biomed. Res. Found., Washington, DC), Vol. 5, Suppl. 3, pp. 353–358.
25. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992) *Comput. Appl. Biosci.* **8**, 275–282.
26. Henikoff, S. & Henikoff, J. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919.
27. Henikoff, S. & Henikoff, J. (1993) *Proteins* **17**, 49–61.
28. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.
29. Staden, R. (1989) *Comput. Appl. Biosci.* **5**, 89–96.
30. Bairoch, A. & Boeckmann, B. (1993) *Nucleic Acids Res.* **21**, 3093–3096.
31. Koonin, E. V. & Dolja, V. V. (1993) *Crit. Rev. Biochem. Mol. Biol.* **28**, 375–430.
32. Gorbalenya, A. E. & Koonin, E. V. (1993) *Curr. Opin. Struct. Biol.* **3**, 419–429.
33. Braithwaite, D. K. & Ito, J. (1993) *Nucleic Acids Res.* **21**, 787–802.
34. Gorbalenya, A. E. & Koonin, E. V. (1990) *J. Mol. Biol.* **213**, 583–591.
35. Gribskov, M. (1992) *Gene* **119**, 107–111.
36. Attwood, T. K. & Findlay, J. B. C. (1993) *Prot. Eng.* **6**, 167–176.
37. Rohde, K. & Bork, P. (1993) *Comput. Appl. Biosci.* **9**, 183–189.
38. Koonin, E. V. (1994) *Nucleic Acids Res.* **22**, 2476–2478.
39. Koonin, E. V., Mushegian, A. R., Tatusov, R. L., Altschul, S. F., Bryant, S. H., Bork, P. & Valencia, A. (1994) *Protein Sci.* **3**, in press.
40. Koonin, E. V. & Tatusov, R. L. (1994) *J. Mol. Biol.*, in press.
41. Alonso, J. C., Stiege, A. C., Dobrinski, B. & Lurz, R. (1993) *J. Biol. Chem.* **268**, 1424–1429.
42. Umezu, K., Chi, N.-W. & Kolodner, R. D. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 3875–3879.
43. Confalonieri, F., Elie, C., Nadal, M., Bouthier de la Tour, C., Forterre, P. & Duguet, M. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 4753–4757.
44. Lovett, S. T. & Kolodner, R. D. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 2627–2631.
45. West, S. C. (1994) *Cell* **76**, 9–15.
46. Lideroth, N. A., Julien, B., Flick, K. E., Calendar, R. & Christie, G. E. (1994) *Virology* **200**, 347–359.
47. Wootton, J. C. & Federhen, S. (1993) *Comput. Chem.* **17**, 149–163.