3836–3845 *Nucleic Acids Research, 1996, Vol. 24, No. 19*

Searching databases of conserved sequence regions by aligning protein multiple-alignments

Shmuel Pietrokovski

Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, WA 98104, USA

Received May 1, 1996; Revised and Accepted August 12, 1996

ABSTRACT

A general searching method for comparing multiple sequence alignments was developed to detect sequence relationships between conserved protein regions. Multiple alignments are treated as sequences of amino acid distributions and aligned by comparing pairs of such distributions. Four different comparison measures were tested and the Pearson correlation coefficient chosen. The method is sensitive, detecting weak sequence relationships between protein families. Relationships are detected beyond the range of conventional sequence database searches, illustrating the potential usefulness of the method. The previously undetected relation between flavoprotein subunits of two oxidoreductase families points to the potential active site in one of the families. The similarity between the bacterial RecA, DnaA and Rad51 protein families reveals a region in DnaA and Rad51 proteins likely to bind and unstack single-stranded DNA. Helix–turn–helix DNA binding domains from diverse proteins are readily detected and shown to be similar to each other. Glycosylasparaginase and gamma-glutamyltransferase enzymes are found to be similar in their proteolytic cleavage sites. The method has been fully implemented on the World Wide Web at URL: http://blocks.fhcrc.org/blocks-bin/LAMA_search

INTRODUCTION

As we acquire more protein sequences and better knowledge of their functions, improved methods are sought to determine sequence similarity. Comparing a protein sequence with a database of protein sequences is currently the standard way to try to identify uncharacterized protein sequences. These types of searches can typically find proteins homologous to the query with sequence similarity levels of ~25% or more (1). The identified proteins usually all belong to one protein family with identical or similar functions in different organisms, cell types and stages of development. Sequences of the family members share common motifs along their entire lengths or in one or more local regions. However, some protein families have members whose sequences have diverged to a point where it is very difficult or impossible to distinguish the sequence similarities due to homology from those due to chance (1).

Sequence similarity between homologous diverged protein sequences can still be detected by comparing multiple alignments of protein families to single sequences (2). The search can be of a single sequence against a database of multiple alignments (3) or of a multiple alignment against a database of sequences (2,4,5). The use of multiple alignments instead of single sequences can increase the sensitivity of the comparison (2,5). Instead of comparing a query sequence with individual sequences, the comparison is between a sequence and a matrix specifying the different amino acids found in each position of the sequences of the recognized family members. This matrix is calculated from a multiple alignment of protein family sequences. Multiple alignments can be across the entire length of the sequences (global) (6) or local—just along the conserved regions (3,7–9). Local alignments of protein families do not demand overall sequence similarity but can still usefully depict the families (10).

This work introduces a searching method for comparing multiple sequence alignments with each other. It is the natural next step following the comparison of sequences with sequences and of sequences with multiple alignments. The method was used to search BLOCKS and other databases of multiply aligned protein sequences (4,9,47). Structural and functional data supported the discovered relationships. Some of the relationships are difficult to detect by other means. The searching method proved to be a practical tool for detecting weak sequence similarities between protein families.

THEORY

Multiple alignments are first transformed into position specific scoring matrices (PSSMs) (2). Each PSSM column corresponds to a sequence position in the multiple alignment. The frequencies of the 20 amino acids (aa) in each position are calculated from the sequences weighted by position-based sequence weights (11). They are then divided by the frequencies expected from protein databases (12). This transformation compensates for over-representation of potentially redundant sequences and for differential background frequencies of the amino acids.

PSSMs are treated as sequences of columns, enabling their alignment with each other. To use algorithms developed for single-sequence alignments (13) we need a measure for comparing pairs of PSSM columns (corresponding to the measures comparing amino acids or nucleotides in single-sequence alignments). The score of the PSSM-to-PSSM alignment is the sum of the scores from comparing the corresponding columns in the two PSSMs (Fig. 1). This procedure enables the searching of multiple alignment databases with multiple alignment queries. Note that comparing the PSSM is separate and independent of its calculation. The PSSMs used to develop and calibrate the method were from local multiple alignments of protein sequences using one possible way for calculating PSSMs. However, the procedure outlined above is general and can be applied to any type of multiple alignment and PSSM.

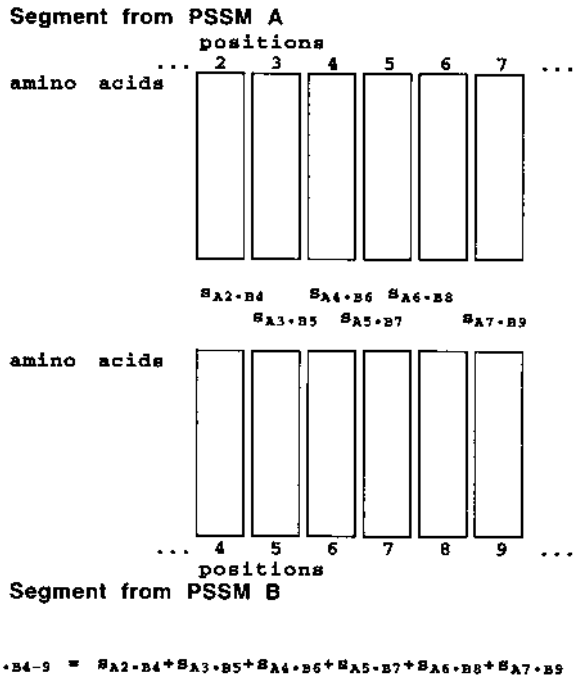

Figure 1. Local alignment of PSSMs. Positions 2–7 from PSSM A aligned with positions 4–9 from PSSM B. A column comparison score, $s_{Xn\cdot Ym}$ is calculated for each pair of positions (A2·B4–A7·B9). The score of the alignment of the two segments, *S*, is the sum of the column comparison scores.

Column comparison measures

Four measures for comparing PSSM columns were tested:
(i) Normalized Euclidean distance (d):

$$d(A,B) = \frac{\sqrt{\sum_{i=1}^{20}(A_i - B_i)^2}}{20}$$

where A_i and B_i are the values of amino acid i in columns A and B, respectively. The distance scores range from 0, for identical columns, to

$$\frac{\sqrt{2 \cdot S^2}}{20}$$

where S is the sum of the values in each column. This largest possible distance occurs when only one different amino acid appears in each column.
(ii) Pearson's correlation coefficient [r,14]:

$$r(A,B) = \frac{\sum_{i=1}^{20}(A_i - \overline{A}) \cdot (B_1 - \overline{B})}{\sqrt{\sum_{i=1}^{20}(A_i - \overline{A})^2 \cdot \sum_{i=1}^{20}(B_1 - \overline{B})^2}}$$

where $\overline{A}$ and $\overline{B}$ are the means of the values in columns A and B, respectively. The correlation scores range from 1, for columns with identical value distributions, to –1 for columns with opposite value distributions (in each column exactly 10 aa occur and those 10 aa are different in the two compared columns).
(iii) Spearmann rank correlation coefficient (rho): this is identical to (ii) except that the ranks of the amino acid values are used instead of the values themselves.
(iv) The sum of the products of the values of corresponding amino acids (p):

$$p(A,B) = \sum_{i=1}^{20} A_i \cdot B_i$$

This measure is an extrapolation of the commonly used measure for scoring a sequence element (amino acid) against a PSSM column (2). The scores of this measure range from S^2, for columns where a single identical amino acid occurs in both columns, to 0 where the columns share no common amino acid.

All of the tested measures compare the sequence weighted amino acid odds ratios of the columns directly, without using a substitution scoring matrix (e.g., PAM, BLOSUM etc.). Substitution scoring matrices were developed for single-sequence alignments that are distinct from the multiple sequence alignments we wish to compare. Unlike single-sequence residues, each column is the result of a multiple sequence alignment and can have different degrees of conservation. However, substitution matrices could be used in calculating the PSSMs (2,5,15).

The column comparison measures were tested on two blocks sets derived from BLOCKS 7.01. Each test set contained blocks from families that had at least 20 sequences in BLOCKS 7.01. To construct a test set, 10 sequences were chosen at random from a BLOCKS 7.01 family and the PROTOMAT block making system (3) applied to the sequences. This was done twice for each BLOCKS 7.01 family, generating two test sets. Test blocks constructed from the same BLOCKS 7.01 family and containing the same sequence regions were considered true positives.

PSSM comparison algorithm

Our method uses the Smith–Waterman algorithm (16) to find optimal local alignments of PSSM pairs. Because the PSSMs are derived from multiple alignments of short ungapped conserved regions, no gaps were allowed in the alignment. The use of a local alignment algorithm allows the detection of partially overlapping blocks:

and of common regions embedded in blocks:

Distribution of chance alignment scores

To determine alignment scores expected by chance and to find a way to compare alignments of different widths, a large number of shuffled blocks were aligned. The BLOCKS 8.0 database was first purged of 215 compositionally biased blocks. Such blocks were defined as having a significant fraction (>12% of all column to column scores within the block) of similar columns [p(A,B) > 0.24)]. Most of these blocks were composed of sequence regions mainly composed of one amino acid or of short simple repeats.

These blocks would be less affected by the shuffling than compositionally unbiased blocks. Each of the remaining 2669 blocks was transformed into a PSSM and the PSSM columns randomly permuted. The resulting PSSMs were compared with the unbiased blocks from the original BLOCKS 8.0 database (excluding comparisons of blocks against their shuffled versions). The best chance alignment score for each of these seven million comparisons was saved.

Implementation

The method has been implemented in a computer program called LAMA (Local Alignment of Multiple-Alignments). The program is written in the C programming language and uses procedures from the BLIMPS program package (17). It is available from the author upon request. The program is integrated into our Blocks World Wide Web site (URL http://blocks.fhcrc.org/blocks-bin/LAMA_search) to search a multiple alignment against databases of blocks and to compare multiple alignments provided by the user. Help files and a number of supporting programs are also included at the site.

RESULTS

Choice of column comparison measure

The column comparison measure is analogous to the amino acid substitution score of single-sequence alignments. However, unlike the 20 discrete amino acids composing protein sequences, multiple-alignment columns can have a range of amino acid distributions. There is no single accepted method for comparing two distributions, so that several possible ones had to be tested empirically.

Four comparison measures were tested for their ability to identify true positive relations between two blocks test sets. All the tested measures gave similarly good results (Table 1). This indicates the robustness of the multiple alignment search method. The method performed well with different column comparison measures. The Pearson correlation coefficient (r) was chosen to be the column comparison measure since it has three properties useful for aligning PSSMs: (i) The difference between scores for similar and for dissimilar columns is natural—similar columns have positive scores and dissimilar columns have negative scores. (ii) Unlike the distance and sum of products measures, the score range of correlation coefficients is independent from the sum of values in the columns and thus does not need any additional normalization. (iii) The distribution of Pearson correlation scores from a very large number of column pairs from the Blocks Database has a negative mean (Fig. 2). A negative mean score is an essential requirement for local optimal alignment algorithms (18). The reason for the negative mean value of the scores is that a Pearson correlation score between conserved columns (where only a few amino acids occur) with different amino acids is slightly negative. The Blocks Database contains multiple alignments of conserved sequence regions where many columns are completely conserved.

Estimating the significance of alignment scores

The mean and the variance of chance alignment scores depend on the length of the alignments since the score of an alignment is the sum of its column scores. Comparison of longer blocks finds longer alignments that have higher scores. A length correction is needed to prevent longer blocks from outscoring equally significant shorter blocks. Grouping the chance alignment scores according to the length of the shorter block gave similar chance score distributions for all the partitions. The number of standard deviations away from the mean score (Z score) within the appropriate partition of the shuffled data was used to estimate the score's significance. This allows comparison of alignments of different lengths and evaluation of chance occurrence.

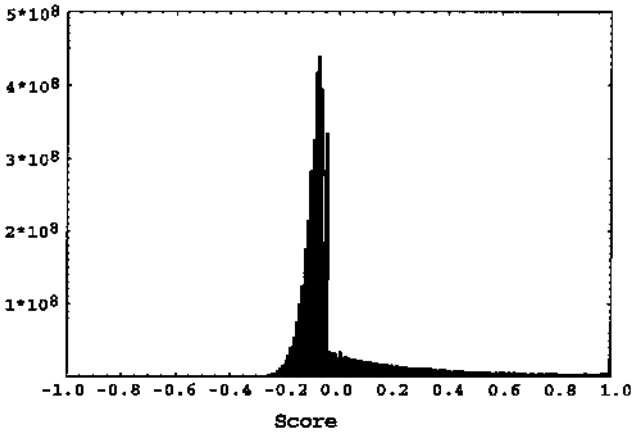

Figure 2. Distribution of column scores from the Blocks Database. The Pearson correlation coefficient was calculated between all column pairs for every pair of blocks from BLOCKS 8.0 (4.48×10^9 scores). The peak at score position –0.05 is due to scores between different wholly-conserved columns (where only one amino acid occurs).

Table 1. Evaluation of column comparison scores[a]

Measure	TNs[b] above 99.995%[c]	TPs below 99.995%[c]	Rank of lowest TP	Equivalence number[d]	ROC area[e]
d	6	29	2467	17	0.9917
r	7	30	5446	15	0.9958
rho	7	30	9413	17	0.9949
p	8	31	2277	19	0.9929

[a]The alignments were global, i.e., the shorter block was 'slid' across the longer block. The total number of scores was 5.4 million and the number of true positives (TPs) 293.
[b]True negatives (TNs) possibly include uncatalogued TPs.
[c]The rank of the 99.995% is 270.
[d]The equivalence number (73) is the number of TPs below the rank where it is equal to the number of TNs above that rank. The number is 0 when all the TPs are above the TNs.
[e]The receiver operating curve (ROC) (74) shows the number of TNs as the x-axis and the number of TPs above that TN as the y-axis. The area under the curve is 1 when all the TPs are above the TNs.

To evaluate the method, each entry in the Blocks Database (3) version 8.6 (3174 blocks from 858 protein families) was searched against the other entries in the database. Every block pair was aligned using the Smith–Waterman algorithm and Pearson correlation coefficient column comparison measure. Cutoff values were chosen that gave no multiple or single hits in the shuffled data: this data had 40% more scores than the evaluation comparisons. All block pairs with Z scores 5.6 and higher (defined as high scores) were saved. A detailed examination was

made of protein families with more than one such similar block pair (multiple block hits), and of protein families with a single block hit with Z score 8.3 or higher. Using these cutoffs resulted in 141 pairs of families (Table 2). Eighty percent of these could be identified as genuine relationships (true positives), 8% were judged to result from similar compositional biases and the rest are unknown.

Several multiple alignment databases, in addition to Blocks, were also searched. Selected relationships below illustrate the usefulness of the method.

Flavoproteins FAD binding and catalytic sites

The succinate dehydrogenase (Sdh) and fumarate reductase (Frd) enzyme complexes contain an FAD flavoprotein subunit. These prokaryotic and eukaryotic enzymes can be grouped into one family by their functional and sequence similarities. The conserved regions of this family include the FAD-binding and the active sites (19). The first site had a very high score (Z score 10.0) with a conserved region from D-amino oxidases FAD flavoenzymes (DAO). This was supported by another hit in a DAO region by the Sdh/Frd active site (Z 5.5).

Table 2. Distribution of top scoring family pairs

Relation type	Genuine[a]	Biased Composition	Unknown	Total
Multiple block hits				
independent[b]	24	–	1	25
repeats[c]	11	6	9	26
inner repeats[d]	15	4	2	21
Single block hits	63	1	5	69
Total	113	11	17	141
Fraction	80%	8%	12%	

[a]Genuine relations were identified by the families Prosite descriptions, by detailed analysis of the literature or by sharing common sequences (22 of the single and independent-multiple hits).
[b]An independent multiple hit is two different protein families related by two or more different high scoring block pairs.
[c]A repeat multiple hit is two different protein families where a block from one family is similar with two or more blocks from the other family.
[d]An inner-repeat multiple hit is a case where the similarities are between blocks from the same family.

The FAD AMP-binding sites in both families are β–α–β ADP binding folds (20) and were already noted as such (19,21). An invariant histidine in the second DAO region (Fig. 3A) was found important for enzymatic activity of porcine DAO (22). This histidine is aligned with a conserved and essential histidine in the Sdh/Frd flavoproteins catalytic site (19,23). Other positions in these aligned regions are similar (column scores 0.31–0.98) as well (Fig. 3B). Note that the dissimilar positions have column scores close to zero (0.04 to –0.14). Our results support the putative identification of the active site in porcine DAO and suggest the position of that site in other members of the DAO family.

BLAST searches (24) of the SwissProt protein database (25) with the protein sequences from each of the two families did not identify sequences from the other family (not shown). Optimal local alignments (16) of all the sequence pairs from the two families had scores expected by chance (not shown). Searching the Blocks Database with the sequences from the two families identified the relation between the families with six Sdh/Frd flavoproteins sequences [multiple hits with 98.1–76.2 percentiles of scores with shuffled sequence queries and P values ranging from 8.4×10^{-3} to 1.1×10^{-1} (10)] but not with the other two sequences from that family or any of the sequences from the DAO family (single hits with less than 60.0 score percentiles).

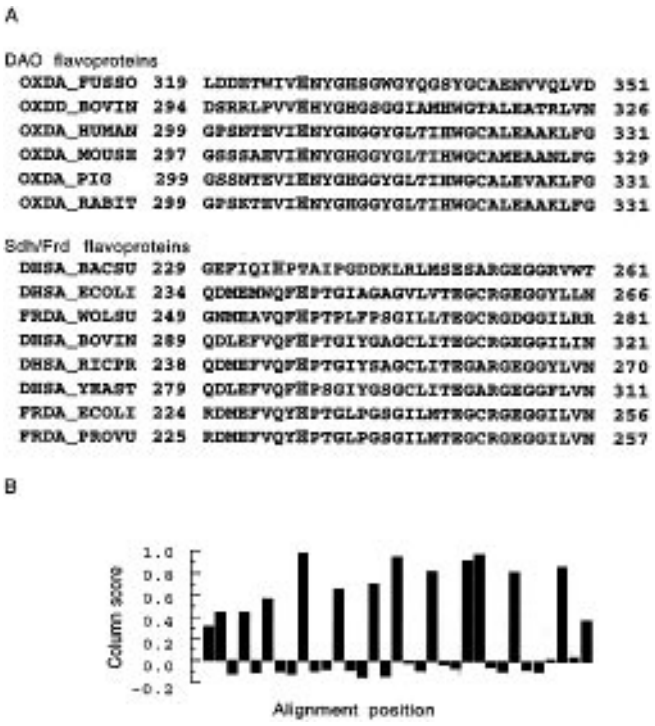

Figure 3. Suggested catalytic site of DAO flavoproteins. (**A**) Positions 17–49 of DAO flavoproteins (block BL00677D) aligned with the catalytic region of Sdh/Frd flavoproteins (positions 3–35 of block BL00504D). The histidines important for the enzymes' catalytic activity are outlined (the histidine in sequence DHSA_BACSU is misaligned due to a two aa insertion). The start and end coordinates flank the sequences. (**B**) The column scores of the alignment.

Single-stranded DNA-binding domains of bacterial-RecA, DnaA and Rad51/Dmc1 proteins

Bacterial RecA proteins are DNA binding proteins mediating various processes including DNA replication, repair and recombination. RecA proteins bind single- and double-stranded DNA, are ATPases and form polymeric filaments that can interact with each other and with DNA (26,27). Bacterial RecA protein sequences are very similar to each other and can be globally aligned (28). Such an alignment is found in the Pfam protein-family alignments database (Sonnhammer EL, personal communication; WWW URL: http://www.sanger.ac.uk/Pfam). Pfam is constructed using hidden Markov models (HMMs) and has long multiple alignments, frequently spanning the entire sequences.

Comparing the multiple alignments in Pfam with the PRINTS (9) database of local multiple alignments identified a multiple hit between the RecA and DnaA protein families. DnaAs are bacterial proteins controlling the initiation of chromosomal

replication. DnaAs cooperatively bind specific DNA sequences, are ATPases regulated by adenine nucleotides, and can melt DNA duplexes (29). The first hit between the RecA and DnaA multiple alignments (Z score 5.6) corresponds to the ATP binding Walker-box A motif (P loop) (30) in both families and is also supported by a weak score (Z 3.7) between the Walker-box B motif of the ATP binding sites (28,31) (Fig. 4A).

The second significant hit (Z 6.0) between the RecA and DnaA families is a 19 aa alignment in their C-terminal half (Fig. 4B). In the *Escherichia coli* RecA crystal structure this region corresponds to the L2 region and the adjacent G α-helix (32). Recently it was shown that a 20 aa peptide, corresponding to *E.coli* RecA L2, can promote homologous DNA pairing by binding and unstacking single-stranded DNA (33). Unstacking of the oriC chromosomal origin results from the binding of DnaA proteins in the priming of DNA replication (34). The region found here to be similar to the RecA DNA binding and unstacking region is likely to perform a comparable function in DnaA proteins.

As far as we know, the relation between the RecA and DnaA protein families was not described previously. BLAST searches of the SwissProt database with individual sequences did not identify any similarity between sequences from the two families (*P* values < 0.9997). Blocks searches of the Blocks Database with the DnaA sequences did not identify the RecA entry (BL00321) containing the regions found in the Pfam entry. Searching the PRINTS database with 61 SwissProt RecA sequences scored hits with the DnaA entry (PR00051) for 17 sequences. Eleven of those had scores higher than expected by chance (10).

Rad51 and Dmc1 proteins were found by functional and sequence similarities to be the eucaryotic and archaebacterial homologs of RecA proteins (35–38). Rad51/Dmc1 sequences from animals, plants and archaebacteria are well conserved and can be aligned in a number of long blocks that can be aligned with the RecA multiple alignment (Fig. 4). The RecA Walker-box A motif and a region preceding it have high scores (Z 5.8 and 9.0) with corresponding regions in the Rad51/Dmc1 proteins. This relation was shown previously (38,39). Two additional weak alignments (Z scores 1.8 and 1.3, expected to appear 0.058 and 0.110 times when comparing two blocks with each other) flank the RecA L2 region (Fig. 4B). Between these Rad51/Dmc1 blocks is a short region, variable among taxa and between Rad51 and Dmc1 proteins. One position is conserved in this region, containing phenylalanine and tyrosine residues (Fig. 5). DNA binding and unstacking activity of *E.coli* RecA were shown to depend on a conserved phenylalanine in the L2 region (33). The Rad51/Dmc1 single-strand DNA binding and unstacking region is likely to be the L2-like region, with the conserved aromatic position crucial for the activity.

Multiple alignment of Rad51/Dmc1 sequences from diverse organisms identified conservation of the ends of the L2-like region and of the aromatic position in its variably-lengthed center. These properties probably follow from the structure of the L2 region. *Escherichia coli* RecA L2 region was found disordered in the crystal when not bound to DNA and to adopt a β-structure upon binding it as a peptide (32,33).

A helix–turn–helix DNA-binding motif in the IS30 family transposases

The excision and insertion of bacterial insertion sequence elements (IS) require the activity of a transposase protein sometimes encoded by the ISs. The IS30 transposase family (40) is represented by five blocks in BLOCKS 8.6. A region of 21 positions from the first block (Fig. 6) had high scores (Z scores 6.7–8.8) only to helix–turn–helix DNA-binding motifs (hth) from four protein families (Fig. 7). Hth DNA binding motifs occur in many proteins that bind specific DNA sequences (41). The IS30 conserved region had high scores with blocks representing hth regions from bacterial regulatory proteins and sigma bacterial transcription initiation factors.

BLAST searches of the SwissProt protein database with the IS30 sequences did not identify any protein with known hth region (not shown). Searching the Blocks Database with the IS30 sequences gave high scores with hth blocks for two of the sequences [98.1 and 93.1 percentiles of scores with shuffled sequence queries (10)]. The other two sequences had low scores with hth blocks (30.8 and 18.1 score percentiles) and higher scores with non-hth blocks. However, each of the transposases putative DNA binding regions was detected by the method of Dodd and Egan (42) as an almost certain hth domain (Fig. 6).

Classification of the first IS30 block as a hth motif is supported by the finding that the N-terminal region of an IS30 transposase, containing the putative hth DNA-binding region, binds the IS30 element (43).

Identifying hth motifs in the Blocks Database

The five hth containing blocks identified in the previous section were used to classify other hth blocks in the Blocks Database. All blocks with high scores (Z score ≥ 5.6) to these hth blocks were examined. Nine more hth blocks were identified in the Blocks Database by having high scoring hits to other hth blocks. The hth blocks were from four types of proteins—bacterial regulatory proteins, homeobox domain proteins, sigma bacterial transcription initiation factors and an IS transposase. All 14 hth blocks had high scoring hits to two or more other hth blocks (Fig. 7). Manual inspection of the Prosite (44) annotation of the protein families in the Blocks Database and of blocks themselves found no other hth blocks in the database. Only two non-hth blocks had high scores to hth blocks. Each was to a single hth block and had a low score relative to the scores between the hth blocks (not shown).

The hth blocks included different numbers of sequences, from 4 to 185. There was no correlation between the number of sequences in a block and its relation to other blocks. This, together with the previous examples, suggests that even blocks with four to six sequences can give a correct representation of conserved protein domains. More than 90% of the blocks in the database used had more than four sequences. This fraction is increasing with each release (>94% in BLOCKS 9.0) as the number of new protein sequences is higher than the number of new protein families (45–47).

Hth regions are a challenging group for sequence comparison methods. This sequence motif is subtle and appears in diverse families that have no other common regions. Hth regions have been used as test groups and incentive for the development of powerful methods for identifying common sequence regions (48–51). The identification of all the hth regions in the Blocks Database shows the potential of the multiple alignment comparison method to aid the annotation of protein-family databases. Besides identifying the function of unknown regions, the approach outlined in this section can be useful in annotating databases that generate the multiple alignments automatically. Multiple alignments of characterized protein motifs (such as the hth, nucleotide

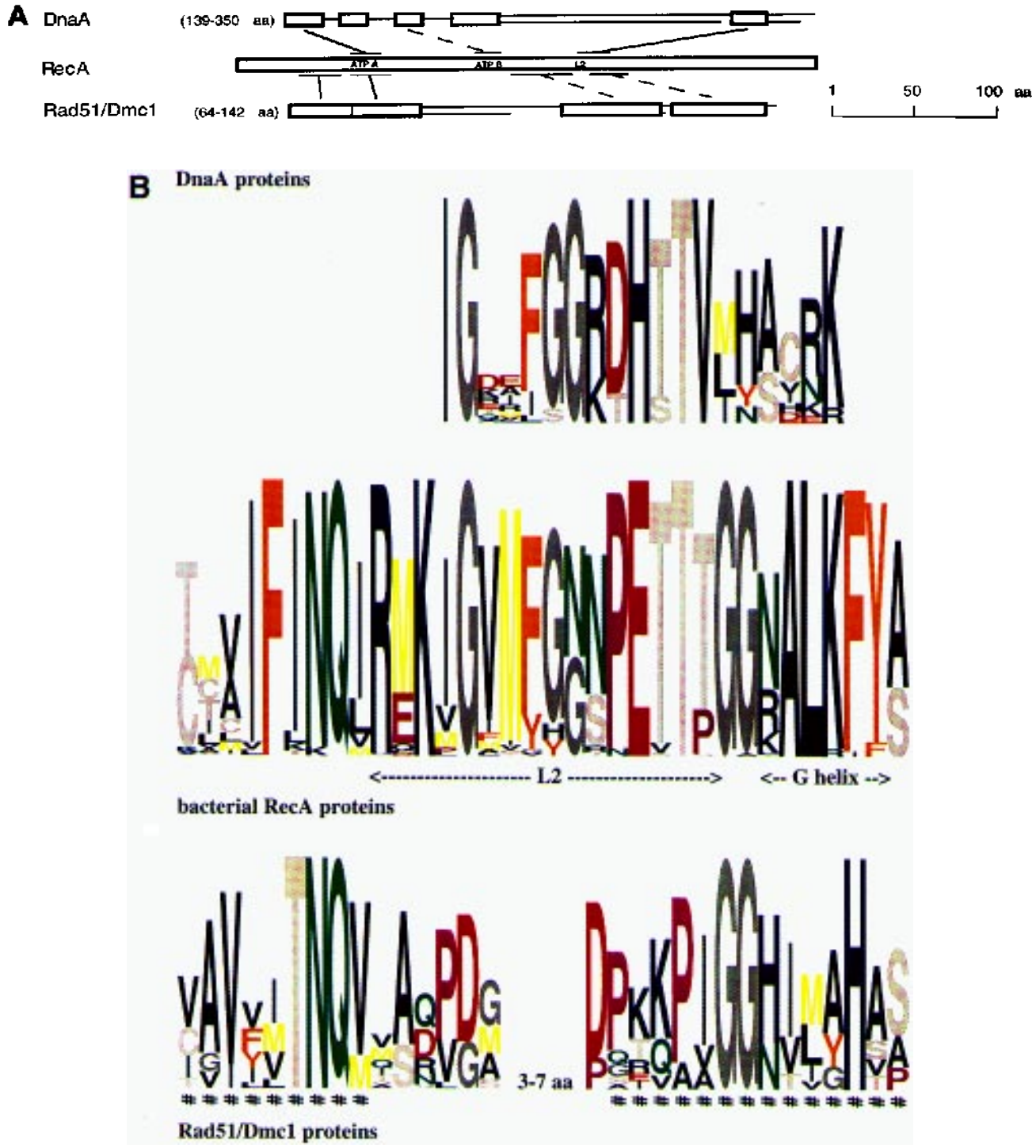

Figure 4. Similarity between the bacterial RecA, DnaA and Rad51/Dmc1 protein families. (**A**) Regions of similarity between the DnaA, Rad51/Dmc1 and RecA multiple alignments. The DnaA (PRINTS entry PR00051 DNAA, 14 sequences) and Rad51/Dmc1 [16 sequences, see Fig. 5, aligned by the BlockMaker program (17)] alignments are composed of blocks shown by boxes. The unaligned sequences between the blocks are indicated by lines. Variable distances between the blocks are shown as two lines corresponding to the shortest and longest distances. The N-terminal unaligned regions are not shown and their lengths are indicated. The RecA alignment (Pfam entry PF00154 recA, 72 sequences) is over all the sequences length, except for a few amino acids at their ends. Lines connect the similar regions of the two families. Weak similarities (see text) are shown by dashed lines. The RecA ATP binding sites A and B and the L2 regions are labeled. (**B**) Alignment of the RecA L2 region with the corresponding DnaA and Rad51/Dmc1 regions. Multiple alignments are shown as sequence logos (72) as described in (17). The amino acids at each position are in the single letter code and are colored according to their chemical and physical properties. The height of each amino acid is proportional to its conservation at that position, after the sequences have been weighted and frequencies adjusted by the expected amino acid composition. The height of the amino acids is directly proportional to their conservation. The positions of the *E.coli* RecA L2 and G α-helix are shown. The DnaA logo is of block PR00051E 1–19 (*E.coli* DnaA 425–443). The RecA logo is of PF00154 194–227 (*E.coli* RecA 187–219). Only the ends of the last two Rad51/Dmc1 blocks and the region in between them are shown (see Fig. 5 for sequence details). The Rad51/Dmc1 positions aligned with RecA are marked by '#'. These alignments extend further and are 36 (left) and 22 (right) aa long.

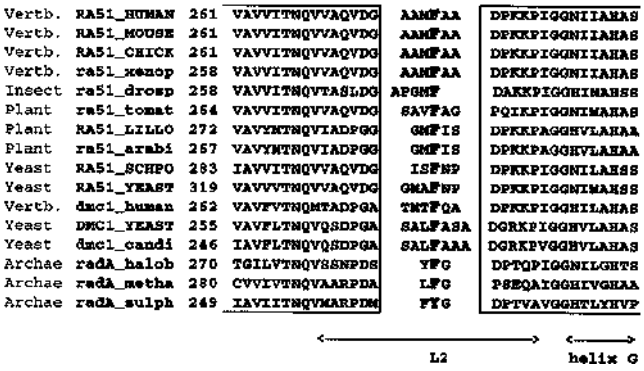

Figure 5. Proposed single-strand DNA binding and unstacking region of Rad51/Dmc1 proteins. Sequences correspond to the Rad51/Dmc1 alignment shown in Figure 4B. Blocks are boxed and the region between them is aligned across a conserved aromatic position (shadowed). The names of the sequences, the start position of the segment shown and their phylogenetic group are shown (vertb. = vertebrate). Underneath the alignment are the regions proposed to correspond with the L2 region and helix G of the *E.coli* RecA crystal structure. The sequence database and accessions are: RA51_HUMAN, SwissProt Q06609; RA51_MOUSE, SwissProt Q08297; RA51_CHICK, SwissProt P37383; ra51_xenop, NCBI g1054624; ra51_drosp, NCBI g693878; ra51_tomat, NCBI g1143810; RA51_LILLO, SwissProt P37384; ra51_arabi, NCBI g1076395; RA51_SCHPO, SwissProt P36601; RA51_YEAST, SwissProt P25454; dmc1_human, NCBI g1321636; DMC1_YEAST, SwissProt P25453; dmc1_candi, NCBI g1145716; radA_halob, NCBI g1378032; radA_metha, NCBI g1378034; radA_sulph, NCBI g1378036.

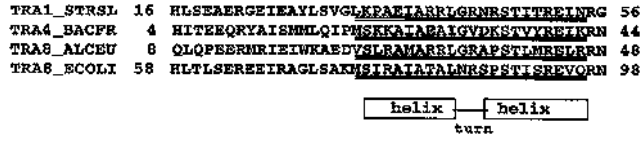

Figure 6. Hth-like region in IS30 transposases. Block BL01043A of the IS30 transposases family. The regions similar to the hth motifs in the block to block searches are underlined. The start and end coordinates flank the sequences. The diagram shows the suggested position of the hth motifs found by the 'hth' algorithm (42). The algorithm scores for hth motifs were 5.19 standard deviation units (SD), corresponding to 100% probability for TRA1_STRSL, 5.95 SD and 100% for TRA4_BACFR, 4.13 SD and 90% for TRA8_ALCEU, and 5.72 SD and 100% for TRA8_ECOLI.

binding folds or leucine zipper) could be used to identify other multiple alignments containing these motifs.

The hth blocks illustrate the problem of distinguishing genuine relationships from chance ones and suggest a solution. Two of the hth blocks (BL00622 and BL01063B) lie below the threshold for detection by single-hit relations (Z score ≥8.3, bold lines in Fig. 7).

Protein families with hth-motifs usually have no other common blocks to support the relation between the hth blocks. However, hth motifs are found in several protein families. These hth blocks all have high scores with each other, but not all of these scores are high enough to identify genuine relationships by themselves. Nevertheless, blocks with a number of such scores to known hth blocks can be identified as hth blocks too (Fig. 7). The two non-hth blocks have high scores to single hth blocks, and do not form part of the connected graph. An analogous strategy is the basis for detecting weak similarities in single-sequence alignments using the BLAST3 program (52).

The similarity shown here between all the hth blocks raises the question how well would one composite hth block identify other hth blocks? A composite hth block, containing 609 sequence segments, is found in the ecmot database (47) constructed by iterative searches using the MOST program (5) (Table 3). Comparing this block with BLOCKS 8.6 identified 18 blocks. Fourteen of those are the hth blocks discussed above (Table 4). All the hth blocks had high to extremely high scores, the lowest one expected to occur 3.2×10^{-3}.

Table 3. PSSM of composite hth block EC0157

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	nulls[b]
A	1	7	5	14	70	6	7	8	1	–	3	3	7	8	3	1	2	2	2	3	3	3	1	7	1
C	1	4	1	3	7	1	2	9	4	9	1	1	1	–	2	–	1	–	5	7	1	–	2	1	4
D	1	4	15	–	–	9	2	–	6	–	5	5	2	–	–	–	1	1	–	2	8	1	4	4	8
E	2	11	29	–	1	19	14	1	5	1	3	11	5	1	–	2	2	1	2	2	4	2	17	9	2
F	8	1	–	1	–	–	2	6	2	5	2	1	–	1	4	7	2	2	29	7	–	5	3	5	5
G	–	4	1	–	3	1	1	–	34	–	2	1	5	1	–	2	1	–	–	2	3	1	6	3	7
H	2	3	4	–	–	6	4	1	11	–	8	2	1	1	3	7	10	14	–	5	2	11	2	4	4
I	16	2	1	40	1	1	4	7	–	18	–	2	1	–	22	1	9	2	16	1	1	5	1	4	3
K	4	11	7	–	–	11	9	–	4	–	2	9	6	1	–	12	6	3	3	17	10	11	5	5	5
L	7	3	1	17	1	2	5	39	–	16	–	2	–	1	14	1	1	2	15	1	1	17	3	2	3
M	7	1	4	5	2	4	9	8	2	14	–	2	5	2	5	1	1	1	6	1	2	6	6	15	1
N	5	4	2	–	–	4	3	–	16	–	6	2	6	2	–	3	6	–	2	9	29	3	4	3	6
P	1	1	–	–	–	–	–	–	–	–	7	13	10	2	–	–	–	–	–	1	–	2	1	2	14
Q	9	7	14	1	2	13	11	–	5	2	2	14	8	24	–	7	7	5	1	24	11	3	6	8	2
R	14	11	7	–	1	15	13	1	2	2	1	20	14	3	1	14	24	3	2	10	14	17	22	6	1
S	–	9	3	–	11	3	4	2	4	1	37	3	15	5	–	19	4	1	–	4	2	1	3	3	4
T	5	12	2	3	1	4	3	6	–	3	20	5	6	.6	0	.	1	0	.	4	4	2	2	5	1
V	11	2	1	15	2	1	1	6	–	21	2	3	2	3	40	–	7	3	10	–	–	1	2	3	4
W	4	2	2	–	–	–	1	3	1	2	–	–	–	–	–	2	3	51	2	–	3	2	1	6	9
Y	2	2	1	–	–	1	5	5	4	6	1	1	2	3	5	17	9	7	2	1	1	6	9	6	2
nulls[b]	2	–	2	11	8	3	1	6	5	7	4	1	3	4	9	4	1	4	5	2	3	1	–	–	86

[a]The PSSM was calculated as described in the text. The position numbers are according to the convention (42,75).
[b]Nulls refer to the number of non occurring amino acids in a row or column.

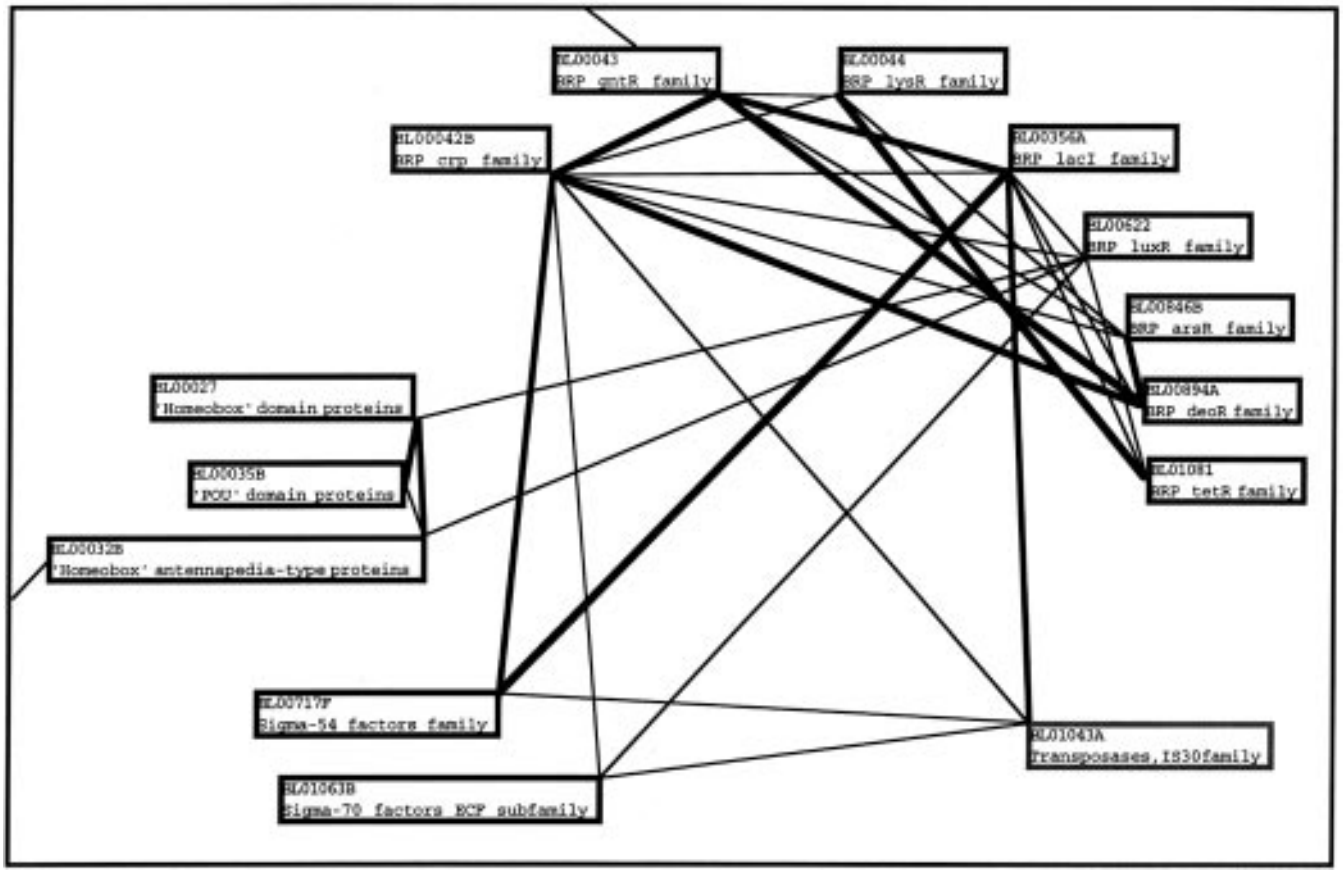

Figure 7. High scores of helix–turn–helix DNA binding blocks. All 14 hth blocks found in BLOCKS 8.6 and their high scoring relationships with each other (true positives) and with other blocks (false positives, outward pointing lines). Each block had different sequences except two pairs of homeobox blocks that had common sequences (BL00027 with BL00032B and with BL00035B). Lines show scores above the 5.6 Z score cutoff. Thick lines correspond to scores above the 8.3 Z score cutoff. BRP, bacterial regulatory proteins.

The four non-hth blocks have significantly lower scores (Z 5.6–6.5) but their similarity can be explained. Two of the blocks are from bacterial regulatory proteins families, occurring C-terminal to the hth motifs. One is a hth-similar region from the araC family (53) and the other corresponds to the hth helix3 and DNA binding hinge helix in the *E.coli lac* repressor protein (54). Another block is from the S3 ribosomal proteins (BL00548A). This protein binds RNA, and it is interesting to note the recent report of the RNA binding activity by a hth domain (55). The last non-hth block is from L-lactate dehydrogenase (LDH) proteins. LDH does not bind DNA but the crystal structure of the detected region (from α2f to βG) is a helix–turn followed by a helix or strand in different proteins (56–58).

The composition and conservation of the composite hth block are clearly related to the structure of the hth motif. Only a few positions are highly conserved but all have some preferred and avoided amino acids (Table 3). For example, proline mainly appears in the turn region and is avoided in the helices while tryptophan is preferred at position 18 and avoided in and around the turn region and in the N-terminal half of the first helix. The last position of the block contains little information and all 20 aa occur in it (Table 3). Nevertheless, methionine is preferred in it much more than the other residues (the value for methionine is about five standard deviations above the mean of the other 19 residues). The PSSM values of the composite hth block are very similar to the hth frequency matrix values found by Dodd and Egan (42). The large amount of diverse sequences and the success of the composite hth block in identifying other hth blocks may offer further insights to the sequence to structure relation of the hth motif.

Proteolytic cleavage sites of glycosylasparaginase and γ-glutamyltransferase enzymes

Glycosylasparaginases (GA) are bacterial, plant and animal enzymes that remove asparagine-linked oligosaccharides from glycoproteins (59). Mutated forms of the enzyme in humans cause aspartylglycosaminuria, a severe disorder of glycoprotein degradation (McKusick 20840). The enzyme is a heterodimer, posttranslationally processed from a single polypeptide (60,61).

The GA family is not present in the current version of the Blocks Database but several different sequences were easily found by keyword and similarity searches of the NCBI non-redundant protein database (National Center for Biotechnology Information). The BlockMaker program (17) identified two blocks in the sequences. Searching version 9.0 of the Blocks Database with the GA blocks found a significant similarity (Z score 7.1, expected occurrence of 0.0095) with a block from the γ-glutamyltransferase (GGT) family (Fig. 8). GGTs catalyzes the transfer of γ-glutamyl moieties to various acceptors (62). Like the GAs, GGTs are heterodimers cleaved from precursor proteins.

Figure 8. Similarity between the glycosylasparaginase and γ-glutamyltransferase protein families. Aligned positions (1–43 in both blocks) of GA and GGT (BL00462F) proteins. The threonines at the C′ side of the cleavage sites are outlined. The start and end coordinates flank the sequences. GA sequence names are the NCBI protein accessions: g555668 *Flavobacterium meningosepticum*; g114276 human; g1213550 *C.elegans*; g231574 *Lupinus arboreus*; g231573 *Lupinus angustifolius*; g496102 *Lupinus albus*; g1076292 *Arabidopsis thaliana*.

Table 4. Blocks similar to composite hth block EC0157

Protein family[a]	Z score
Homeobox' domain proteins	18.4
Homeobox' antennapedia-type proteins	13.2
POU' domain proteins	11.7
BRP crp family	12.1
BRP gntR family	12.4
BRP lysR family	14.4
BRP lacI family	11.5
BRP luxR family	12.5
BRP arsR family	8.0
BRP tetR family	14.1
BRP deoR family	8.7
Sigma-54 factors family	7.8
Sigma-70 factors ECF subfamily	8.3
Transposases, IS30 family	11.2
BRP araC family	6.5
BRP lacI family[b]	6.6
Ribosomal S3 proteins	5.8
L-lactate dehydrogenase family	5.8

[a]The non-hth blocks are at the end of the table separated by a bold line. The family Blocks Database entry numbers are shown in Figure 7 except for BRP araC family, BL00041; L-lactate dehydrogenase, BL00064D; and ribosomal protein S3 proteins, BL00548A.
[b]This hit is with block BL00356B that follows the lacI hth block BL00356A, see text.

An invariable threonine is found at the first position of the block alignment. After cleavage it forms the N-terminal end of the β and small subunits of GAs and GGTs, respectively. This residue is essential for the processing of precursor proteins in both families (63,64) and is the active site of GAs (65). The sequence similarity between the families is over a 43 aa region but was not detected by BLAST and Blocks searches with the individual sequences.

Together with the sequence similarity, the structure of proteins from both families and the corresponding locations of the similar blocks indicate a genuine relationship between the GA and GGT families. This relationship probably indicates similar post-translational processing and maybe some similarity between the active sites (76), that are found at the processing site of GAs and at the small GGT subunits (65,66). A bacterial GA was found to be processed by a novel autoproteolytic reaction, apparently involving a histidine residue two amino acids upstream of the threonine cleavage site (64). This histidine is conserved in the known animal GAs (61) (but not in the plant ones). GGTs do not have a histidine residue at a corresponding position. However, an invariant histidine, two amino acids downstream of the threonine cleavage site (Fig. 8), was found essential for processing of *E.coli* GGT (63). The GGT processing protease is only partially characterized (67) but the light chain of GGT was shown to have proteolytic activity, especially towards the heavy GGT chain (66). Thus the GGT processing protease might be GGT itself.

DISCUSSION

Sequence alignment searching methods proceeded from single-sequence alignments, to aligning sequences with multiple alignments and, now, aligning multiple alignments with multiple alignments. Previously multiple alignment comparison has been used as a step in finding global multiple alignments (68,69) and for visual (dot plot) comparison of profiles (70). The method and scores assessment presented in this paper extend multiple alignment comparison to database searching. The procedure detects similarities both between protein families and between conserved domains in an automated way. This paper demonstrates that in at least some cases our method can recognize relationships not identified by conventional database searches. The sensitivity of the method is due to the comparison between multiple sequences, while its selectivity is due to the use of only conserved sequence regions (regions that can be multiply aligned confidently).

The examples presented in this paper illustrate the practical use of the block to block comparison method. Both local and global multiple alignments (represented here by the Blocks, PRINTS, ecmot and Pfam databases) can be compared by the method. The single-stranded DNA-binding domains and glycosylasparaginase and γ-glutamyltransferase examples show how a search query (multiple alignment) can be made for a group of related protein sequences. Users can search using multiple alignments and sequences not available in public databases. Biological information, such as known structure or functional domains, can then be used to choose or construct a proper multiple alignment thus integrating specific knowledge with primary sequence information.

Currently most of the newly determined protein sequences are already part of some known family (45–47) and many contain known folds (71). Consequently, the strategy described in this paper will become more generally applicable with time. Our method augments the existing tools for detecting related protein sequences and should be used with other available methods for searching and comparing sequences.

ACKNOWLEDGEMENTS

I thank Steven and Jorja Henikoff for many helpful discussions and suggestions and Gary Stormo for critically reading the manuscript. Part of this work was done at the 1995 'Patterns in Biological Sequences' workshop of the Aspen Center for Physics. The author is a Howard Hughes Medical Institute Fellow of the Life Sciences Research Foundation. This work was supported in part by a grant from the NIH to Steven Henikoff (GM29009).

REFERENCES

1 Doolittle, R. F. (1987) *Of URFs and ORFs: a Primer on How to Analyze Derived Amino Acid Sequences.* Oxford University Press, Oxford, UK.
2 Gribskov, M., McLachlan, A. D. and Eisenberg, D. (1987) *Proc. Natl. Acad. Sci. USA*, **84,** 4355–4358.
3 Henikoff, S. and Henikoff, J. G. (1991) *Nucleic Acid. Res.*, **19,** 6565–6572.
4 Henikoff, S. (1992) *New Biol.*, **4,** 382–388.
5 Tatusov, R. L., Altschul, S. F. and Koonin, E. V. (1994) *Proc. Natl. Acad. Sci. USA*, **91,** 12091–12095.
6 Sander, C. and Schneider, R. (1991) *Proteins*, **9,** 56–68.
7 Smith, R. F. and Smith, T. F. (1990) *Proc. Natl. Acad. Sci. USA*, **87,** 118–122.
8 Sonnhammer, E. L. and Kahn, D. (1994) *Protein Sci.*, **3,** 482–492.
9 Attwood, T. K., Beck, M. E., Bleasby, A. J. and Parry-Smith, D. J. (1994) *Nucleic Acid. Res.*, **22,** 3590–3596.
10 Henikoff, S. and Henikoff, J. G. (1994) *Genomics*, **19,** 97–107.
11 Henikoff, S. and Henikoff, J. G. (1994) *J. Mol. Biol.*, **243,** 574–578.
12 Henikoff, S. and Henikoff, J. G. (1996) *Methods Enzymol.*, **266,** 88–105.
13 Pearson, W. R. and Miller, W. (1992) *Methods Enzymol.*, **210,** 575–601.
14 Pearson, K. and Lee, A. (1903) *Biometrika*, **2,** 357–462.
15 Henikoff, J. G. and Henikoff, S. (1996) *Comput. Appl. Biosci.*, **9,** 135–143.
16 Smith, T. F. and Waterman, M. S. (1981) *J. Mol. Biol.*, **147,** 195–197.
17 Henikoff, S., Henikoff, J. G., Alford, W. J. and Pietrokovski, S. (1995) *Gene*, **163,** 17–26.
18 Altschul, S. F. (1991) *J. Mol. Biol.*, **219,** 555–565.
19 Birch-Machin, M. A., Farnsworth, L., Ackrell, B. A., Cochran, B., Jackson, S., Bindoff, L. A., Aitken, A., Diamond, A. G. and Turnbull, D. M. (1992) *J. Biol. Chem.*, **267,** 11553–11558.
20 Wierenga, R. K., Terpstra, P. and Hol, W. G. (1986) *J. Mol. Biol.*, **187,** 101–107.
21 Schulz, G. E., Schirmer, R. H. and Pai, E. F. (1982) *J. Mol. Biol.*, **160,** 287–308.
22 Miyano, M., Fukui, K., Watanabe, F., Takahashi, S., Tada, M., Kanashiro, M. and Miyake, Y. (1991) *J. Biochem.*, **109,** 171–177.
23 Schroder, I., Gunsalus, R. P., Ackrell, B. A., Cochran, B. and Cecchini, G. (1991) *J. Biol. Chem.*, **266,** 13572–13579.
24 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990) *J. Mol. Biol.*, **215,** 403–410.
25 Bairoch, A. and Boeckmann, B. (1991) *Nucleic Acid. Res.*, **19,** 2247–2249.
26 Clark, A. J. and Sandler, S. J. (1994) *Crit. Rev. Microbiol.*, **20,** 125–142.
27 Roca, A. I. and Cox, M. M. (1990) *Crit. Rev. Biochem. Mol. Biol.*, **25,** 415–456.
28 Karlin, S. and Brocchieri, L. (1996) *J. Bacteriol.*, **178,** 1881–1894.
29 Skarstad, K. and Boye, E. (1994) *Biochim. Biophys. Acta*, **1217,** 111–130.
30 Walker, J., Saraste, M., Runswick, M. and Gay, N. (1982) *EMBO J.*, **1,** 945–951.
31 Koonin, E. (1993) *Nucleic Acids Res.*, **21,** 2541–2547.
32 Story, R. M. and Steitz, T. A. (1992) *Nature*, **355,** 374–376.
33 Voloshin, O. N., Wang, L. and Camerini-Otero, R. D. (1996) *Science*, **272,** 868–872.
34 Bramhill, D. and Kornberg, A. (1988) *Cell*, **52,** 743–755.
35 Bishop, D., Park, D., Xu, L. and Kleckner, N. (1992) *Cell*, **69,** 439–456.
36 Shinohara, A., Ogawa, H. and Ogawa, T. (1992) *Cell*, **69,** 457–470.
37 Ogawa, T., Yu, X., Shinohara, A. and Egelman, E. (1993) *Science*, **259,** 1896–1899.
38 Sandler, S. J., Satin, L. H., Samra, H. S. and Clark, A. J. (1996) *Nucleic Acids Res.*, **24,** 2125–2132.
39 Story, R., Bishop, D., Kleckner, N. and Steitz, T. (1993) *Science*, **259,** 1892–1896.
40 Dong, Q., Sadouk, A., van der Lelie, D., Taghavi, S., Ferhat, A., Nuyten, J. M., Borremans, B., Mergeay, M. and Toussaint, A. (1992) *J. Bacteriol.*, **174,** 8133–8138.
41 Pabo, C. O. and Sauer, R. T. (1992) *Annu. Rev. Biochem.*, **61,** 1053–1095.
42 Dodd, I. B. and Egan, J. B. (1990) *Nucleic Acids Res.*, **18,** 5019–5026.
43 Stalder, R., Caspers, P., Olasz, F. and Arber, W. (1990) *J. Biol. Chem.*, **265,** 3757–3762.
44 Bairoch, A. (1991) *Nucleic Acids Res.*, **19,** 2241–2245.
45 Green, P., Lipman, D., Hillier, L., Waterston, R., States, D. and Claverie, J. M. (1993) *Science*, **259,** 1711–1716.
46 Koonin, E. V., Bork, P. and Sander, C. (1994) *EMBO J.*, **13,** 493–503.
47 Koonin, E., Tatusov, R. and Rudd, K. (1995) *Proc. Natl. Acad. Sci. USA*, **92,** 11921–11925.
48 Dodd, I. B. and Egan, J. B. (1987) *J. Mol. Biol.*, **194,** 557–564.
49 Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. and Wootton, J. C. (1993) *Science*, **262,** 208–214.
50 Neuwald, A. F. and Green, P. (1994) *J. Mol. Biol.*, **239,** 698–712.
51 Neuwald, A. F., Liu, J. S. and Lawrence, C. E. (1995) *Protein Sci.*, **4,** 1618–1632.
52 Altschul, S. F. and Lipman, D. J. (1990) *Proc. Natl. Acad. Sci. USA*, **87,** 5509–5513.
53 Brunelle, A. and Schleif, R. (1989) *J. Mol. Biol.*, **209,** 607–622.
54 Lewis, M., Chang, G., Horton, N. C., Kercher, M. A., Pace, H. C., Schumacher, M. A., Brennan, R. G. and Lu, P. (1996) *Science*, **271,** 1247–1254.
55 Dubnau, J. and Struhl, G. (1996) *Nature*, **379,** 694–699.
56 Grau, U., Trommer, W. and Rossmann, M. (1981) *J. Mol. Biol.*, **151,** 289–307.
57 Abad-Zapatero, C., Griffith, J., Sussman, J. and Rossmann, M. (1987) *J. Mol. Biol.*, **198,** 445–467.
58 Iwata, S. and Ohta, T. (1993) *J. Mol. Biol.*, **230,** 21–27.
59 Mononen, I., Fisher, K., Kaartinen, V. and Aronson, N. J. (1993) *FASEB J.*, **7,** 1247–1256.
60 Fisher, K., Tollersrud, O. and Aronson, N. J. (1990) *FEBS Lett.*, **276,** 232.
61 Tarentino, A., Quinones, G., Hauer, C., Changchien, L. and Plummer, T. J. (1995) *Arch. Biochem. Biophys.*, **316,** 399–406.
62 Lieberman, M., Barrios, R., Carter, B., Habib, G., Lebovitz, R., Rajagopalan, S., Sepulveda, A., Shi, Z. and Wan, D. (1995) *Am. J. Pathol.*, **147,** 1175–1185.
63 Hashimoto, W., Suzuki, H., Yamamoto, K. and Kumagai, H. (1995) *J. Biochem. (Tokyo)*, **118,** 75–80.
64 Guan, C., Cui, T., Rao, V., Liao, W., Benner, J., Lin, C. and Comb, D. (1996) *J. Biol. Chem.*, **271,** 1732–1737.
65 Kaartinen, V., Williams, J., Tomich, J., Yates, J. d., Hood, L. and Mononen, I. (1991) *J. Biol. Chem.*, **266,** 5860–5869.
66 Gardell, S. and Tate, S. (1979) *J. Biol. Chem.*, **254,** 4942–4945.
67 Kuno, T., Matsuda, Y. and Katunuma, N. (1984) *Biochem. Int.*, **8,** 581–588.
68 Taylor, W. R. (1988) *J. Mol. Evol.*, **28,** 161–169.
69 Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) *Nucleic Acids Res.*, **22,** 4673–4680.
70 Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) *Comput. Appl. Biosci.*, **10,** 19–29.
71 Orengo, C. (1994) *Curr. Opin. Struct. Biol.*, **4,** 429–440.
72 Schneider, T. D. and Stephens, R. M. (1990) *Nucleic Acids Res.*, **18,** 6097–6100.
73 Pearson, W. R. (1995) *Protein Sci.*, **4,** 1145–1160.
74 Metz, C. E. (1978) *Sem. Nuclear Med.*, **8,** 283–298.
75 Pabo, C. and Sauer, R. (1984) *Annu. Rev. Biochem.*, **53,** 293–321.
76 Brannigan, J., Dodson, G., Duggleby, H., Moody, P., Smith, J., Tomchick, D. and Murzin, A. (1995) *Nature*, **378**, 916–919.