PROTEINS: Structure, Function, and Genetics 9:180–190 (1991)

# A Workbench for Multiple Alignment Construction and Analysis

**Gregory D. Schuler, Stephen F. Altschul, and David J. Lipman**
*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894*

***ABSTRACT*** **Multiple sequence alignment can be a useful technique for studying molecular evolution, as well as for analyzing relationships between structure or function and primary sequence. We have developed for this purpose an interactive program, MACAW (Multiple Alignment Construction and Analysis Workbench), that allows the user to construct multiple alignments by locating, analyzing, editing, and combining "blocks" of aligned sequence segments. MACAW incorporates several novel features. (1) Regions of local similarity are located by a new search algorithm that avoids many of the limitations of previous techniques. (2) The statistical significance of blocks of similarity is evaluated using a recently developed mathematical theory. (3) Candidate blocks may be evaluated for potential inclusion in a multiple alignment using a variety of visualization tools. (4) A user interface permits each block to be edited by moving its boundaries or by eliminating particular segments, and blocks may be linked to form a composite multiple alignment. No completely automatic program is likely to deal effectively with all the complexities of the multiple alignment problem; by combining a powerful similarity search algorithm with flexible editing, analysis and display tools, MACAW allows the alignment strategy to be tailored to the problem at hand.**

**Key words: pattern recognition, sequence alignment, algorithms, amino acid sequences, molecular sequence data, proteins, software**

## INTRODUCTION

The working biologist often is confronted with a set of protein sequences that he thinks may share some similar function or structure. He may not be certain a priori whether this similarity is shared by all the proteins he is considering, or by only a subset. Indeed, the case may be more complicated, with the set of proteins dividing into two or more related groups, but with the groups exhibiting no clear similarity to one another. The similarity among the proteins may be confined to certain small regions, and these regions may appear near the amino terminus of some of the proteins but near the carboxyl terminus of others. Without any sure knowledge as to which of these possibilities might hold, the biologist would like a tool to draw his attention to any regions of similarity shared by any subset of the proteins under consideration.

When all the sequences under study are known to be globally related, various automatic multiple alignment strategies are effective.[1–13] However in the general case, some of the input sequences may be fragmentary, some may share isolated regions of similarity with one another, and some may be unrelated to the rest. The central difficulty is that a simple definition of the problem frequently fails to encompass the complex relationships possible among multiple sequences. Automatic methods can seek local similarities, but absent an intelligent program some judgment is usually required to integrate the information generated. Multiple sequence editors have therefore been developed to aid the researcher.[14] The usefulness of such editors is of course limited by the power and generality of the similarity search algorithms they incorporate.

The question of how to find regions of local similarity among multiple sequences is not trivial. Several attacks on the problem have been mounted,[6,15–20] but each has suffered from certain limitations, among which have been the necessity of specifying the length of a putative pattern beforehand, restrictions on the allowable length or position of such a pattern, and the requirement or preference that any pattern found appear in all the sequences tested. In this paper we describe an interactive program MACAW (Multiple Alignment Construction and Analysis Workbench), that allows the user to locate, analyze, assess the statistical significance of, edit, and combine regions of local similarity among multiple protein sequences. MACAW implements a new strategy for finding similar regions

Received April 23, 1990; revision accepted July 11, 1990.
Address reprint and software requests to Dr. Gregory D. Schuler, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894.

among multiple sequences that is free from most of the limitations of earlier methods. These regions serve as the fodder for interactive editing analysis.

One advantage of multiple over pairwise alignments is that while a pattern found in only two sequences may not appear significant, the same pattern found in many sequences may nevertheless be surprising. A local multiple alignment program should therefore be sensitive to weak, or statistically insignificant, pairwise similarities. There is little to be gained, however, in reporting multiple alignments in which certain pairs of sequences exhibit no detectable similarity whatsoever. The search strategy of MACAW is based upon these observations.

The main obstacle to finding regions of similarity common to many sequences is the immense size of the space that must be searched. Given $n$ protein sequences each of length $l$, a region of similarity common to all may begin by aligning any one of approximately $l^n$ different sets of amino acids. Searching a space this large is totally impractical for typical $l$ on the order of 100, and $n$ greater than three. The central idea of MACAW is to impose a single condition on the alignments it seeks: that all segments show a minimal amount of mutual similarity. This one condition allows us to eliminate virtually the entire space of possible alignments in $O(n^2l^2)$ time. The remaining alignments can then be carefully analyzed and their statistical significance assessed. This basis approach has been used elsewhere to search protein databases, with a single query sequence, for multiple alignments.[21]

When seeking multiple as opposed to pairwise alignments, many novel problems arise, not only of an algorithmic, but also of a definitional and statistical nature. We discuss a variety of such issues in this paper, and describe the features of MACAW that allow the user to deal with them.

## TERMINOLOGY AND OUTLINE

Given a set of $n$ sequences, imagine choosing one segment of some specific length from each of $m$ of the sequences. We will call such a set of segments an *m-block*, or simply a *block*. A block imposes upon the sequences from which its segments are chosen a specific phase or alignment. We will call any set of $m$ sequences locked into a specific alignment, with no gaps allowed, an *m-diagonal* or simply a *diagonal*. Since a diagonal need not comprise all $n$ sequences under study, one diagonal may contain or be contained in another.

The program MACAW consists of several subroutines of use in constructing multiple alignments. The central search routine identifies diagonals that may contain a block of interest. Several methods are provided for parsing a diagonal into blocks that may represent significant similarities among the sequences. The statistical significance of any block can be tested using recently described statistical results.[22,23] Blocks may be edited and combined in any consistent manner to produce an alignment of segments from some or all of the input sequences. A future possibility is to incorporate into MACAW a program for global multiple alignment[10–12] to allow regions between blocks, which may contain gaps, to be optimally aligned. We discuss below the main features of each of these procedures, and give some examples of the program's use.

## THE COST OF PAIRWISE AND MULTIPLE ALIGNMENTS

In order to choose among possible alignments, one needs some measure of alignment quality. The usual procedure when comparing two protein sequences is to choose a set of *similarity scores* for aligning pairs of amino acids, and seek an alignment that maximizes the total score.[24–29] If gaps are to be allowed, scores are also assigned to gaps of various lengths.[30–35] The search routine of MACAW, which seeks diagonals containing interesting blocks, requires only scores for aligning amino acids. The most widely used set of such scores is the PAM-250 matrix[36] which was derived from a study of amino acid replacements in homologous proteins; a variation of this matrix is used by MACAW.

There are a variety of ways in which similarity scores may be generalized to the alignment of amino acids from more than two sequences. An aligned set of $n$ amino acids we call an *n-column* or simply a *column*. We require a way to assign a score to any column; the score of a block is then simply the sum of the scores assigned to each of its columns. One simple rule is to take the score of a column to be the sum of all the pairwise similarity scores of the amino acids it comprises;[4,17] we call these *SP-scores* for "Sum of the Pairs." Given some knowledge of an evolutionary tree relating the sequences in question, a more biologically realistic set of scores can be defined.[1,2,37] Because we assume we have no such knowledge a priori, MACAW uses SP-scores, but the program can easily be modified to use other scoring schemes. As described below, MACAW employs multiple alignment scores to estimate the statistical significance of interesting blocks.

## MAXIMALLY CONSISTENT DIAGONALS

The central idea of MACAW's search routine is to seek only blocks in which all pairs of segments are contained in pairwise subalignments with score greater than or equal to some threshold value. This makes biological sense, because reporting a block should imply that all the segments it aligns are related. If a given pair of segments in a block evince not even the slightest relationship, then it makes little sense to consider that block. Because relationships may become apparent from the comparison of many sequences that are not yet apparent from the

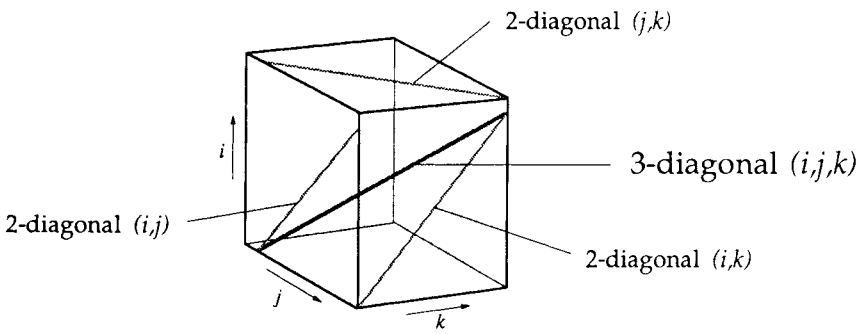



Fig. 1. A 3-diagonal and the three 2-diagonals it contains.

comparison of just two, the threshold for accepting possible pairwise relationships should be as low as possible.

The search routine begins by comparing all pairs of input sequences. For each pair it locates those diagonals that contain a 2-block with score at least $T$. The threshold $T$ may be chosen so that about 10% of the diagonals for even unrelated sequences are marked; with such a low threshold, no detectable relationship should be missed. For $n$ sequences of total (aggregate) length $L$, the time complexity of this step is $O(L^2)$ because each sequence must be compared with all others. Generally this is the most time-consuming step of MACAW.

Having recorded which 2-diagonals show potential relationship, the search procedure attempts to extend each such diagonal to one containing more sequences. The only requirement is that for any multiple diagonal formed, all implied 2-diagonals must have been marked during the original step. Thus a diagonal containing three sequences is formed only when each of the three 2-diagonals it contains has been marked (Fig. 1). A diagonal is reported only if there is no way in which to extend it further.

A quick calculation shows why the number of diagonals reported generally does not mushroom with an increasing number $n$ of input sequences. Consider an arbitrary diagonal of $m$ sequences. Each of the $\binom{m}{2}$ 2-diagonals it contains has approximately a 10% chance of having been marked. Most pairwise diagonals will have been marked because of some chance similarity, so that these probabilities are essentially independent. The probability this diagonal will be reported is therefore about $(0.1)^{\binom{m}{2}}$. If each input sequence has length $l$, calculation shows that the number of $m$-diagonals is approximately $m\binom{n}{m}l^{m-1}$. While this number grows exponentially with $m$, the probability that a given $m$-diagonal is reported decreases exponentially with $m^2$. If $n$ or $l$ are large, the threshold $T$ may be adjusted upward to decrease the probability that a random 2-diagonal is marked. For a typical run, virtually never is a "random diagonal" of more than four sequences reported. Since diagonals are generated only by extending those that have already been found, this stage of MACAW runs very quickly.

While random diagonals are quickly excluded with increasing $m$, there is a case that causes MACAW some problems. This is when many of the input sequences contain related internal repeats. The number of diagonals containing blocks that represent true relationships can then grow exponentially with $m$. By using only a portion of each repetitive sequence, it is easy to prevent this from happening.

## PARSING A DIAGONAL

We have described how MACAW identifies diagonals of potential interest. Given such a diagonal, we are then faced with the question of what block or blocks to report. By analogy to pairwise sequence comparison, this may seem trivial: given an $m$-diagonal, simply find the $m$-block with maximum similarity score. Some reflection shows that this answer is too glib.

If we know beforehand that any homology that exists within a diagonal must be shared by all the sequences it includes, and must be shared over the same stretch, then maximizing the similarity score of an $m$-block is reasonable. However, much more complicated situations can present themselves. We illustrate in Figure 2 two simple situations that can arise in a diagonal containing only three sequences: bold lines represent regions of the sequences that are actually homologous. In Figure 2a, nowhere do all three sequences share a homology; the diagonal has nevertheless been reported because each pair of aligned sequences contains a high scoring block. Depending upon the actual amino acids in each sequence, the highest scoring 3-block might overlap one, all or none of the actual pairwise homologies; in no case, however, does it accurately represent the true relationships among the sequences. In Figure 2b, all three sequences are homologous for a short stretch, after which only two continue to be homologous. The highest scoring 3-block would most likely contain the whole region homologous for all three


Fig. 2. **(a)** Pairwise similarities in a 3-diagonal that have no mutual agreement. **(b)** Pairwise similarities in a 3-diagonal that agree but vary in extent.

sequences. Depending upon how strong the remaining pairwise homology is, the block might contain all, none, or part of it. In the first instance, a three-sequence homology would be claimed over a stretch where none exists; in the second, a significant pairwise relationship would be missed; in the third, both these would occur. Ideally, we would like our program to be sensitive to these sorts of complexities. The basic problem is that the homologies that exist in an $n$-diagonal may best be represented by $n$-blocks in some regions, 4-, 3-, or 2-blocks in others, and a combination of different, disconnected blocks in yet others. It is difficult to capture this richness by maximizing a single well-defined score.

The approach MACAW takes is heuristic, but has proved effective. It has the effect of parsing the diagonal blocks containing various numbers of segments; each block represents a possible homology among the segments it includes. As discussed later, the statistical significance of any block can be assessed knowing its score and the number of segments it contains. The procedure MACAW uses for parsing a multiple diagonal has two steps.

Step 1. For each 2-diagonal contained in a given diagonal, MACAW locates its highest scoring 2-block. The score of this block must be at least $T$, and MACAW marks all amino acid pairs it contains. This block excluded, if the highest scoring 2-block that remains also has score at least $T$, MACAW marks the amino acid pairs it contains, and so forth. This allows for the recognition two or more separate regions of similarity. Marking a pair of amino acids signifies there is some evidence they are homologous; this information is used in the following step.

Step 2. MACAW turns now to considering individual columns of the multiple diagonal. A given column aligns amino acids from all or some of the sequences the diagonal contains. We represent the column by a graph (Fig. 3): each amino acid is a vertex, and two vertices are connected by an edge if the pair of amino acids they represent was marked in step 1. The graph may have no edges, in which case there is no evidence that any of the amino acids in the column are homologous. However if the graph does contain edges, we group into a block those vertices that fall into connected components. In graph $G_4$ of Figure 3 for example, vertices $b$, $c$, and $d$ form one block, vertices $a$ and $e$ another, while vertex $f$ is not placed into a block. A block represents possible homology between the amino acids it contains. After examining each column MACAW may have formed many blocks, but each only one column long. To simplify matters, whenever blocks from adjacent columns involve exactly the same sequences, MACAW coalesces them into a single block, as shown with graphs $G_2$ and $G_3$ of Figure 3. These are the blocks that are finally reported. As shown in Figure 3, when displaying a given diagonal, MACAW colors each residue to indicate the number of edges incident to it in the graph representing its column. This gives the user some additional idea of which regions of a diagonal or block reflect the greatest mutual similarity among the sequences.

While this parsing procedure is able to disentangle some of the complex relationships described above, it does not always choose the boundaries of blocks ideally. MACAW therefore provides the user several tools for editing blocks. To illustrate the relationships in a block more clearly, MACAW colors each amino acid according to the number of edges incident to its corresponding vertex (as described in step 1 and Fig. 3 above). The user may remove a sequence from a block or adjust its left and right boundaries. He may also automatically find the boundaries of a block that maximize its score. As discussed earlier, this maneuver can be deceptive, but it is nevertheless frequently useful.

## STATISTICAL SIGNIFICANCE

Given a specific block, an important question is whether it represents some real relationship, or whether it can be explained simply by chance. This question is a thorny one, with many issues arising for the comparison of multiple sequences that do not arise for the comparison of just two. We discuss some of these issues here.

To decide whether a block is statistically significant, one needs to have a model of chance. We can model a single protein by assuming that at any position each type of amino acid has a specific probability of occurring. We assume these probabilities are position independent and have no Markov dependence. While this is a simplified model for real proteins, it at least gives us a handle on the ques-

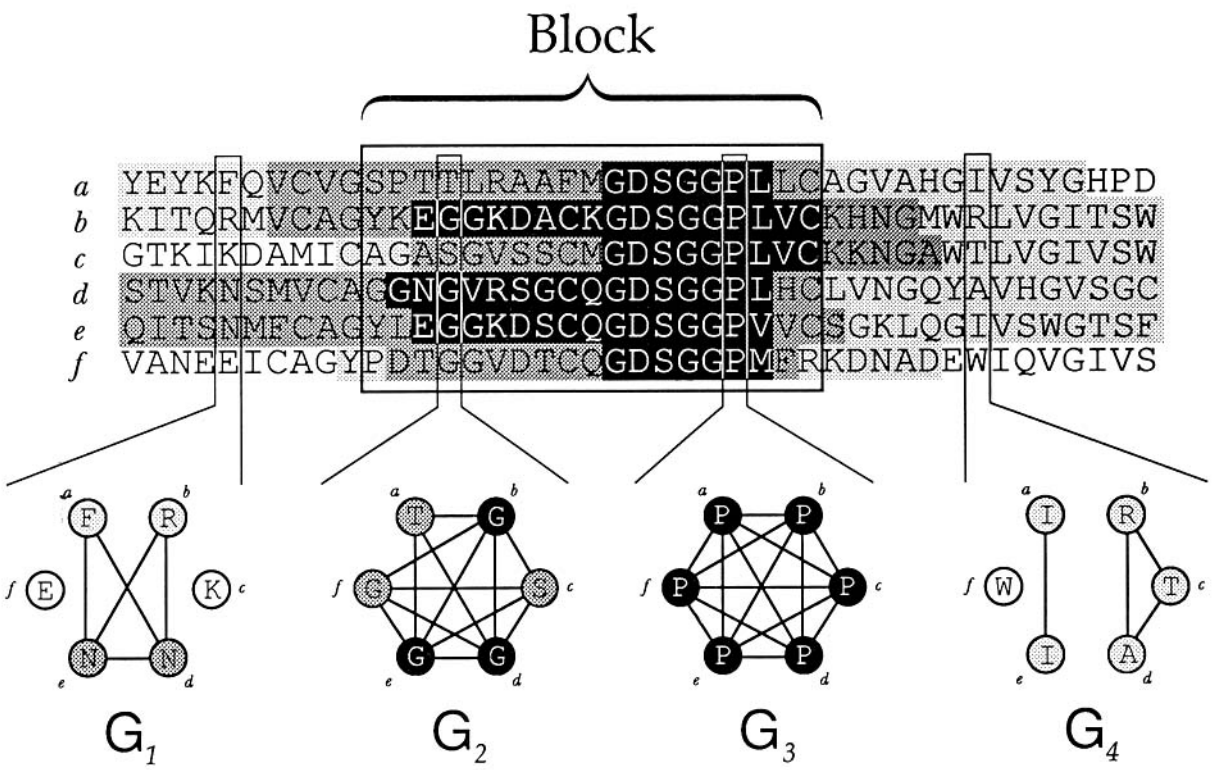



Fig. 3. Parsing a diagonal. Each column is represented by a graph showing putative relationships among the residues. Adjacent columns whose graphs connect residues from the same sequences are coalesced into blocks. Each residue is colored to indicate the number of edges it touches in the graph representing its column.

tion. Now consider a set of $n$ unrelated proteins of lengths $l_1, \ldots, l_n$. Any possible column aligning amino acids from the $n$ sequences then has an associated probability, which is the product of the probabilities for the amino acids it contains. It also has an associated score, which in our case is the SP-score. What is the highest score for an $n$-block that can be expected to occur by chance? New statistical results answer just this question.[22,23] In brief, the statistical significance of an $n$-block with score $S$, when found by searching sequences of lengths $l_1, \ldots, l_n$, is given by the formula

$$1 - \exp(-KN\, e^{-\lambda S}) \qquad (1)$$

where $N$ is the product of the $l_i$, and $K$ and $\lambda$ are constants determined by the possible scores for an $n$-column and their corresponding probabilities.[22,23] This formula presupposes that a column's score is constant under permutation of the column's elements, that the expected score for a column is negative, and that the lengths of the $n$ sequences being compared are not too dissimilar. All these are valid assumptions for the comparison of $n$ unrelated protein sequences of typical size using SP-scores.

Suppose that we have found an $n$-block $B$ whose score is too high to be readily explainable by chance under the above model; what might this mean? It is of course possible that our random protein model is at fault. The high score of $B$ may simply reflect some Markov dependence of amino acid residues in real proteins, or possibly unusual amino acid usage in the specific proteins under consideration. Imagine, however, that the high score of $B$ reflects some true relationship. Can we conclude that all the segments that comprise $B$ are mutually related? The answer is no, for the null hypothesis of nonrelatedness can be violated by just a subset of the segments in $B$. Even two closely related segments can boost the score for $B$ to such an extent that the whole block can appear statistically significant. Any similarity the remaining sequences show may be readily explainable by chance. The case may of course be even more complicated, for a block may divide into two or more subsets of segments, each of which is internally related but whose mutual similarity is due to chance. The heuristic procedure described above for parsing diagonals seeks to avoid reporting blocks of this sort, but we have no direct way of deciding which among the segments of a significant block are mutually related. Nevertheless, this problem may be addressed by removing from the block any segments that seem only marginally similar to the rest. If the statistical significance of the block is unaffected or increases as a result, the similarity of the remaining segments to

those that were removed may be ascribable to chance.

Finally, how do we assess the significance of an $m$-block found from a comparison of $n$-sequences, where $m < n$? For example, we may have input six protein sequences, and found a high-scoring 4-block $B$. Had we compared just the four sequences whose segments contribute to $B$, formula (1) would apply, but we have in effect performed a much larger search. While rigorous results concerning the significance of $B$ remain to be proved, analogy with a related problem[38] provides a tentative solution. We simply set $N$ in formula (1) to the sum, over all possible choices of $m$ sequences, of the product of the sequence lengths: this is the effective size of the space searched. The parameters $K$ and $\lambda$ are of course those for $m$-sequence comparison. MACAW uses this procedure to estimate the statistical significance of any $m$-block. The numbers produced are based, of course, on a random protein model that can only approximate the sequence structure of real proteins, and therefore they must be used only as a rough guide.

## EXAMPLE: CONSTRUCTION OF A MULTIPLE ALIGNMENT

Imagine that a molecular biologist has newly acquired sequence data for some proteins about which little is known, but which may possibly be related to a family of well-studied proteins. There is much to be gained by attempting to align these sequences with known and unknown functions, since the discovery of mutual similarities could shed new light on the biological properties of the unknown proteins. For the sake of example, we will select three sequences of the serine protease family of enzymes bovine trypsin, bovine chymotrypsin, and pig elastase to represent the well-studied proteins. To this set, let us add three "unknown" proteins, the first of which is a distantly related serine protease, *Streptomyces griseus* trypsin. In order to represent the common situation in which newly generated data are fragmentary, we will also use the sequence of mast cell proteinase II (another serine protease) which has had the first 60 amino acids (37% of the protein) removed. Finally, we will include a human globin sequence, which is completely unrelated to the serine proteases. Once the input sequences have been loaded into the MACAW working environment, the first step in the alignment construction process is to search for blocks of similarity. The user has control over two parameters which affect the outcome of this search—the pairwise score threshold $T$, and the minimum number of sequences $m_{min}$ that a block must contain to be reported. As a starting point, it is typical to use a relatively high score threshold, $T = 35$ in this example, so that the most highly similar regions can be quickly identified, leaving a more exhaustive search until later in the analysis. Since we have three sequences that are known to be related and three unknowns, we will set $m_{min} = 4$, so that every block reported will include at least one of the unknown sequences. Using these parameters, MACAW finds 11 blocks, which are listed in Figure 4. For each block, the number of sequences $m$, the length of the block $l$, the score $S$, and the statistical significance $P$ are given. In addition, Figure 4 shows more detailed views for three of the blocks (numbers 1, 4, and 6). A "global view" of the block shows its constituent sequence segments highlighted on a schematic diagram of the alignment in order to give the user an indication of its spatial layout. In a "local view," the block is marked by a box on a display of the diagonal from which it was extracted. In other words, the sequences are offset such that the sequence segments are brought into alignment. Superimposed on the sequence text is information on the local interrelatedness of the sequences, encoded in the form of color (as described above and in the legend to Fig. 3). Taking together all of the information generated by MACAW, the user can judge the relevance of each block.

It is noteworthy that none of the blocks reported by MACAW in this example includes all six of the input sequences. This is clearly due to the inclusion of the globin sequence ("unknown 3"), which is not represented in any of the blocks reported, even those with low scores. Based on this observation, the user may conclude that the sequence is unrelated to the others and remove it from MACAW's operating environment. However this is not essential, as its presence will not interfere with further analysis and construction of the alignment.

In many cases, it may be possible to select a block that is "better" than the one generated by MACAW. For example, the block shown in Figure 5a contains a region of high similarity on the left (indicated by the dark background), but its right half is somewhat less similar (lighter backgrounds). In Figure 5b, the block boundaries have been edited to include only the region of strongest similarity. As a consequence of this change, the score of the block increased from 389 to 423. It is also possible for the user to remove any sequence from the block that is deemed unrelated to the others. Once a block has been located and (optionally) edited, it can be "linked" into the alignment. Figure 6 illustrates the linking process, which involves the insertion of gaps into the sequences such that the segments comprising the block come into alignment. Moreover, linked residues will remain aligned despite subsequent changes elsewhere in the alignment. For example, the linking of the block (black region) in Figure 6 does not disturb downstream blocks that have been previously linked (light gray boxed in the diagram and upper case text in the sequence). Visual cues on the display allow the user to easily distinguish regions that have been aligned from those still requir-

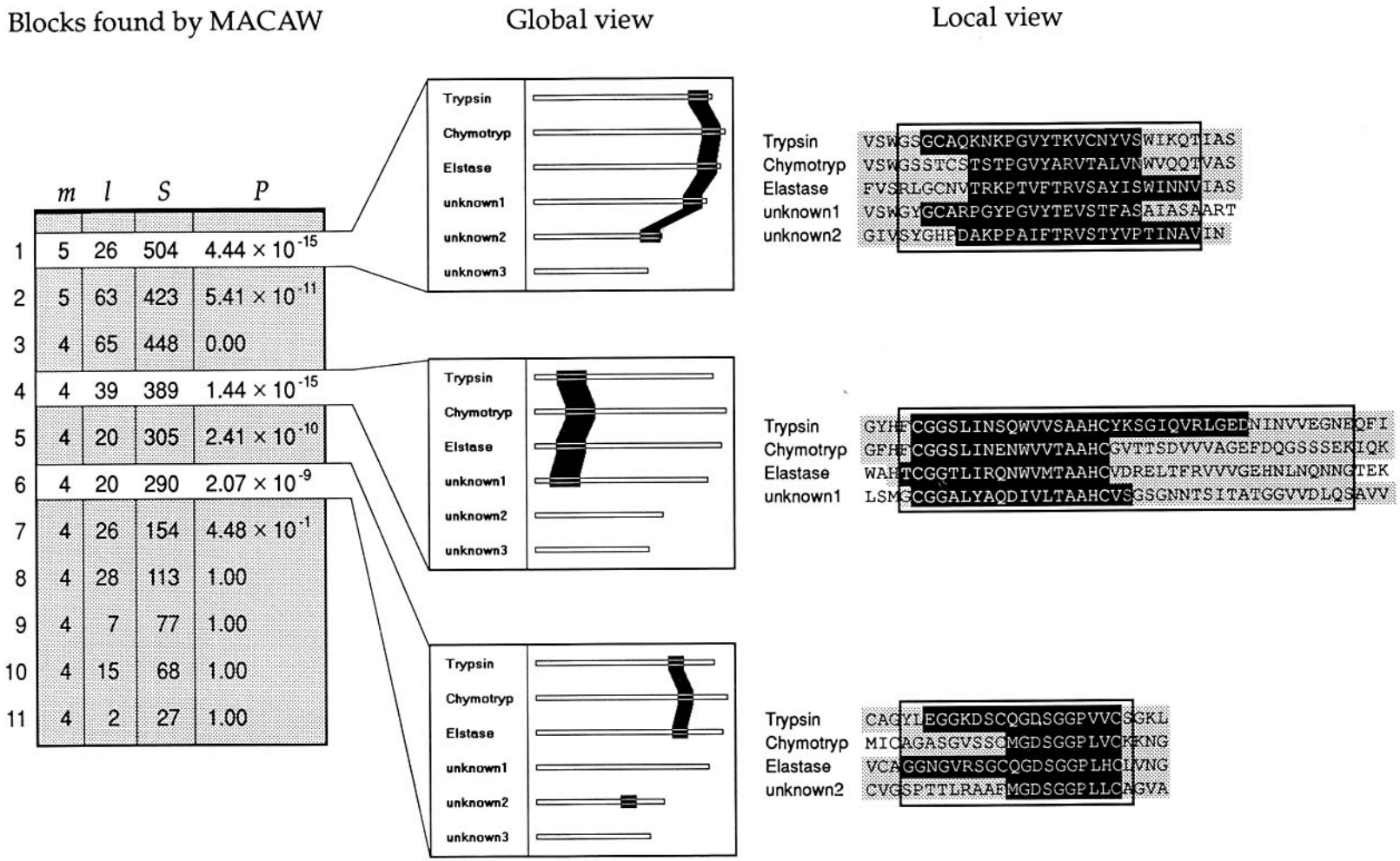



Fig. 4. Three ways in which MACAW reports blocks: (i) in tabular form; (ii) a global view representing the position of each segment in its respective sequence; (iii) a local view showing the alignment of the residues.

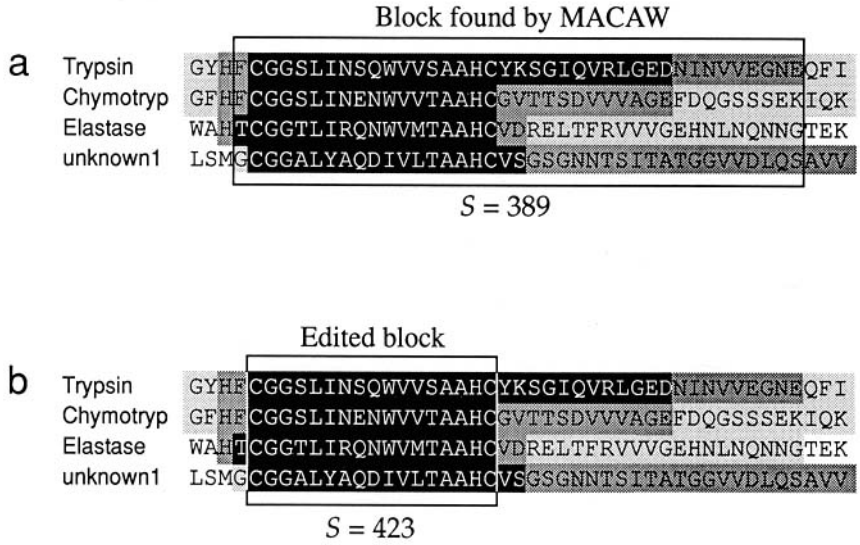



Fig. 5. Editing an individual block.

ing attention. The user may later choose to "unlink" any linked region as desired. After linking all of the significant blocks located with $T = 35$, further analysis using a lower threshold, $T = 20$ in this example, can be directed toward unaligned regions in between the linked blocks. A final pass with $T = 10$ was used to produce the alignment shown in Figure 7, which is in almost complete agreement with an alignment of the trypsin, chymotrypsin and elastase sequences based on three-dimensional structure.[39] This divide-and-conquer strategy often yields the desired result in the least amount of time. Although additional analysis can be performed with a lower score threshold, we have generally found this to be less fruitful, as the blocks reported are predominantly low scoring and not statistically significant. It is also possible to edit the alignment manually, if desired. In many cases, however, biologists will be satisfied with an incomplete alignment since it will emphasize the regions of highest similarity, hence greater potential biological interest.

For the sake of flexibility, MACAW provides an alternate method by which blocks may be located, which is based on a user-supplied sequence pattern. This strategy makes use of a common regular expression matching algorithm to locate the occurrences of the pattern in each of the sequences. In a subsequent step, a list of $m$-blocks is generated by enumerating all possible combinations of $m$ sequence segments, where m is greater than or equal to the $m_{\min}$ value entered by the user. For example, a block similar to (but shorter than) the first block shown in Figure 4 could have been generated by searching for the pattern "P?[VI][FY]??V," where the question mark character may match any symbol and square brackets enclose alternate symbols for a particular position. A particular pattern can either be chosen to reflect prior knowledge about the problem, or can be abstracted from blocks found by using the search strategy described above. The ability to look for these sorts of patterns can be quite useful in certain situations. Basing local alignments on such patterns (though not predefined) is basically the approach described elsewhere.[19,20]

## CONCLUSION

The problem of analyzing multiple sequences for mutual relationships is a complex one, and it seems unlikely that a single definition or algorithm can capture its richness. The problem can hardly be addressed, however, without the aid of the computer. We believe that MACAW, by integrating a fast and powerful similarity search algorithm with varied multiple alignment editing and analysis procedures, provides a flexible new tool for molecular biologists studying relationships among multiple protein and DNA sequences.

The MACAW executable program is available from the authors upon request. While it can load up to 16 sequences, its performance is most satisfactory with eight or fewer. The program was developed for the Microsoft® Windows™ (version 3.0 or later) graphic user interface, which imposes the following system requirements: a personal computer running the MS-DOS® operating system (version 3.1 or later), at least 640K of memory, and a graphics display. In addition, MACAW requires a mouse or other pointing device and additional memory will significantly improve its performance (2M is recommended).

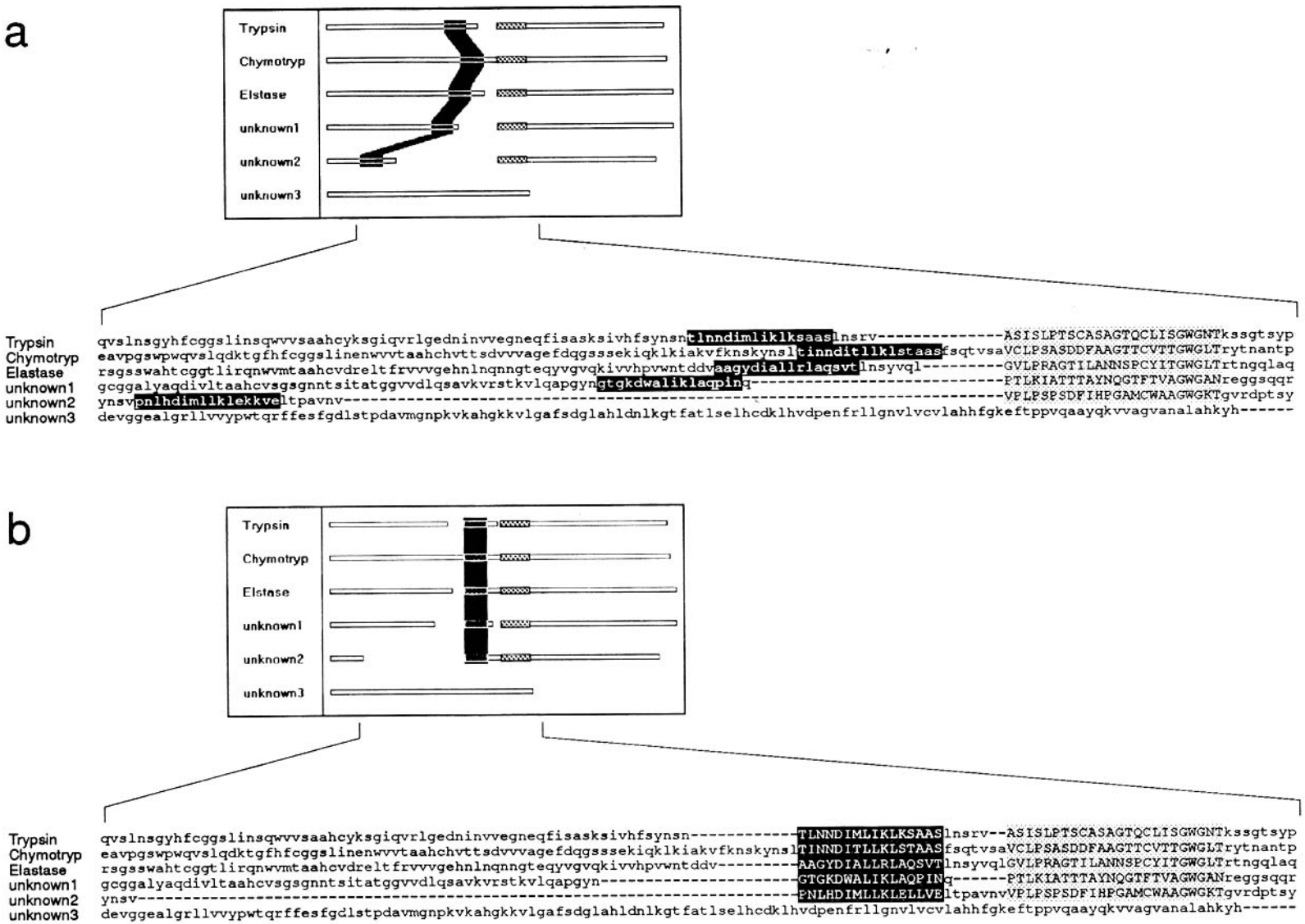
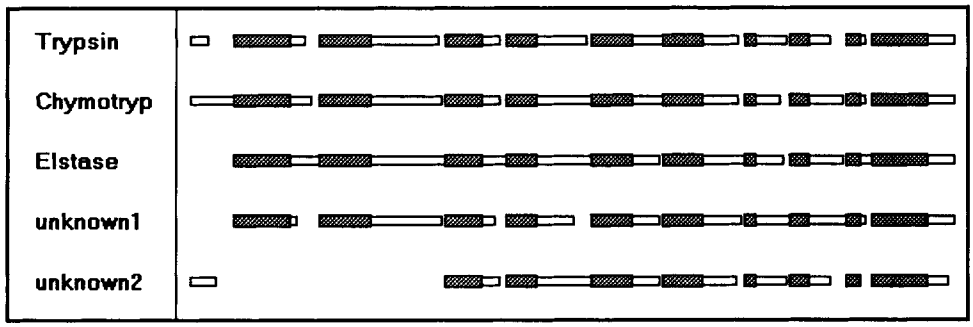



Fig. 6. Linking blocks into an alignment. **(a)** A new block under consideration. **(b)** The block linked into alignment, along with a previously linked block.

a




b

```
Trypsin    vddddk---------IVGGYTCGANTVPYQVSLNsgyhf-----CGGSLINSQWVVSAAHCyksgi  52
Chymotryp  cgvpaiqpvlsglsrIVNGEEAVPGSWPWQVSLQdktgfhf---CGGSLINENWVVTAAHCgvtts  63
Elastase   ---------------VVGGTEAQRNSWPSQISLQyrsgsshahtCGGTLIRQNWVMTAAHCvdrel  51
unknown1   ---------------VVGGTRAAQGEFPFMVRLSmg--------CGGALYAQDIVLTAAHCvsgsg  43
unknown2   rkrestqqk-------------------------------------------------------  9

Trypsin    qvrlgedninvvegnwqf--ISASKSIVHPSYNsntlnn--DIMLIKLKSAaslnsrvasislpts  114
Chymotryp  dvvvagefdqgsssekiqk-LKIAKVFKNSKYNsltinn--DITLLKLSTAasfsqtvsavclpsa  126
Elastase   tfrvvvgehnlnqnngteqyVGVQKIVVHPYWNtddvaagyDIALLRLAQSvtlnsyvqlgvlpra  117
unknown1   nntsitatggvvdlqsavk-VRSTKVLQAPGYNgtgk----DWALIKLAQPinqptlkiattta--  102
unknown2   --------------------IKVEKQIIHESYNsvpnlh--DIMLLKLEKKveltpavnvvplpsp  53

Trypsin    ca--SAGTQCLISGWGNTkssgtsypdvLKCLKAPILSDSSCksaypgqitsnm--FCAGyleggk  176
Chymotryp  sddfAAGTTCVTTGWGLTrytnantpdrLQQASLPLLSNTNCkkywgtkikdam--ICAGasgvss  190
Elastase   gtilANNSPCYITGWGLTrtngqlaqt-LQQAYLPTVDYAICssssywgstvknsmVCAGgngvrs  182
unknown1   ----YNQGTFTVAGWGANreggsqqry-LLKANVPFVSDAACrsaygnelvanee-ICAGypdtgg  162
unknown2   sdfiHPGAMCWAAGWGKTgvrdptsyt-LREVELRIMDEKACvdyryyeykfq---VCVGspttlr  115

Trypsin    dscq-GDSGGPVvcsgklq-----GIVSWgs--GCAQKNKPGVYTKVCNYVSwikqtiasn      229
Chymotryp  cm---GDSGGPLvckkngawtlv-GIVSWgs--STCSTSTPGVYARVTALVNwvqqtlaan      245
Elastase   gcq--GDSGGPLhclvnggyavh-GVTSFvsrlGCNVTRKPTVFTRVSAYISwinnviasn      240
unknown1   vdtcqGDSGGPMfrkdnadewiqvGIVSWgy--GCARPGYPGVYTEVSTFASaiasaartl      221
unknown2   aafm-GDSGGPLlcagvah-----GIVSY----GHPDAKPPAIFTRVSTYVPtinavin       164
```

Fig. 7. A multiple alignment produced by MACAW. **(a)** Schematic global view. **(b)** Local view with aligned residues capitalized and highlighted.

## REFERENCES


1. Sankoff, D. Minimum mutation trees of sequences. SIAM J. Appl. Math. 28:35–42, 1975.
2. Sankoff, D., Cedergren, R.J. Simultaneous comparison of three or more sequences related by a tree. In: "Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison." Sankoff, D., Kruskal, J.B., eds. Reading, MA: Addison-Wesley, 1983: 253–263.
3. Waterman, M.S., Perlwitz, M.D. Line geometries for sequence comparisons. Bull. Math. Biol. 46:567–577, 1984.
4. Murata, M., Richardson, J.S., Sussman, J.L. Simultaneous comparison of three protein sequences. Proc. Natl. Acad. Sci. U.S.A. 82:3073–3077, 1985.
5. Johnson, M.S., Doolittle, R.F. A method for the simultaneous alignment of three or more amino acid sequences. J. Mol. Evol. 23:267–278, 1986.
6. Sobel, E., Martinez, H. A multiple sequence alignment program. Nucl. Acids Res. 14:363–374, 1986.
7. Barton, G.J., Sternberg, M.J. A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. J. Mol. Biol. 198:327–337, 1987.
8. Feng, D., Doolittle, R.F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. 25:351–360, 1987.
9. Taylor, W.R. Multiple sequence alignment by a pairwise algorithm. Comput. Appl. Biosci. 3:81–87, 1987.
10. Carrillo, H., Lipman, D. The multiple sequence alignment problem in biology. SIAM J. Appl. Math. 48:1073–1082, 1988.
11. Altschul, S.F., Lipman, D.J. Trees, stars, and multiple biological sequence alignment. SIAM J. Appl. Math. 49:197–209, 1989.
12. Lipman, D.J., Altschul, S.F., Kececioglu, J.D. A tool for multiple sequence alignment. Proc. Natl. Acad. Sci. U.S.A. 86:4412–4415, 1989.
13. Vingron, M., Argos, P. A fast and sensitive multiple sequence alignment algorithm. Comput. Appl. Biosci. 5:115–121, 1989.
14. Barber, A.M., Maizel, J.V., Jr. SequenceEditingAligner: A multiple sequence editor and aligner. Gene Anal. Techn. Appl. 7:39–45, 1990.
15. Queen, C.M., Wegman, N., Korn, L.J. Improvements to a program for DNA analysis: A procedure to find homologies among many sequences. Nucl. Acids Res. 10:449–456, 1982.
16. Waterman, M.S., Arratia, R., Galas, D.J. Pattern recognition in several sequences: consensus and alignment. Bull. Math. Biol. 46:515–527, 1984.

17. Bacon, D.J., Anderson, W.F. Multiple sequence alignment. J. Mol. Biol. 191:153–161, 1986.
18. Stormo, G.D., Hartzell, G.W., III Identifying protein-binding sites from unaligned DNA fragments. Proc. Natl. Acad. Sci. U.S.A. 86:1183–1187, 1989.
19. Posfai, J., Bhagwat, A.S., Posfai, G., Roberts, R.J. Predictive motifs derived from cytosine methyltransferases. Nucl. Acids Res. 17:2421–2435, 1989.
20. Smith, H.O., Annau, T.M., Chandrasegaran, S. Finding sequence motifs in groups of functionally related proteins. Proc. Natl. Acad. Sci. U.S.A. 87:826–830, 1990.
21. Altschul, S.F., Lipman, D.J. Protein database searches for multiple alignments. Proc. Natl. Acad. Sci. U.S.A. 87: 5509–5513, 1990.
22. Karlin, S., Altschul, S.F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. U.S.A. 87: 2264–2268, 1990.
23. Karlin, S., Dembo, A., Kawabata, T. Statistical composition of high-scoring segments from molecular sequences. Ann. Stat. 18:571–581, 1990.
24. Needleman, S.B., Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol. 48:443–453, 1970.
25. Sellers, P.H. On the theory and computation of evolutionary distances. SIAM J. Appl. Math. 26:787–793, 1974.
26. Smith, T.F., Waterman, M.S. Comparison of biosequences. Adv. Appl. Math. 2:482–489, 1981.
27. Sankoff, D., Kruskal, J.B. "Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison." Reading, MA: Addison-Wesley, 1983.
28. Waterman, M.S. General methods of sequence comparison. Bull. Math. Biol. 46:473–500, 1984.
29. Sellers, P.H. Pattern recognition in genetic sequences by mismatch density. Bull. Math. Biol. 46:501–514, 1984.
30. Waterman, M.S., Smith, T.F., Beyer, W.A. Some biological sequence metrics. Adv. Math. 20:367–387, 1976.
31. Gotoh, O. An improved algorithm for matching biological sequences. J. Mol. Biol. 162:705–708, 1982.
32. Fitch, W.M., Smith, T.F. Optimal sequence alignments. Proc. Natl. Acad. Sci. U.S.A. 80:1382–1386, 1983.
33. Altschul, S.F., Erickson, B.W. Optimal sequence alignment using affine gap costs. Bull. Math. Biol. 48:603–616, 1986.
34. Miller, W., Myers, E.W. Sequence comparison with concave weighting functions. Bull. Math. Biol. 50:97–120, 1988.
35. Myers, E.W., Miller, W. Optimal alignments in linear space. Comput. Appl. Biosci. 4:11–17, 1988.
36. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. A model of evolutionary change in proteins. In "Atlas of Protein Sequence and Structure," Vol. 5, Suppl. 3. Dayhoff, M.O., ed. Washington: Natl. Biomed. Res. Found., 1978: 345–352.
37. Altschul, S.F., Carroll, R.J., Lipman, D.J. Weights for data related by a tree. J. Mol. Biol. 207:647–653, 1989.
38. Karlin, S., Ghandour, G. Comparative statistics for DNA and protein sequences: single sequence analysis. Proc. Natl. Acad. Sci. U.S.A. 82:5800–5804, 1985.
39. Greer, J. Comparative model-building of the mammalian serine proteases. J. Mol. Biol. 153:1027–1042, 1981.