RESEARCH ARTICLE

Rapid and Sensitive Protein Similarity Searches

David J. Lipman and William R. Pearson

Abstract. *An algorithm was developed which facilitates the search for similarities between newly determined amino acid sequences and sequences already available in databases. Because of the algorithm's efficiency on many microcomputers, sensitive protein database searches may now become a routine procedure for molecular biologists. The method efficiently identifies regions of similar sequence and then scores the aligned identical and differing residues in those regions by means of an amino acid replaceability matrix. This matrix increases sensitivity by giving high scores to those amino acid replacements which occur frequently in evolution. The algorithm has been implemented in a computer program designed to search protein databases very rapidly. For example, comparison of a 200-amino-acid sequence to the 500,000 residues in the National Biomedical Research Foundation library would take less than 2 minutes on a minicomputer, and less than 10 minutes on a microcomputer (IBM PC).*

Technical advances in molecular biology are providing scientists with the primary sequences of proteins implicated in such critical biological processes as differentiation and transformation. Frequently, the only information available about a potentially interesting protein is its amino acid sequence. This information has become more useful because of the availability of large protein-sequence databases and, increasingly, relationships have been discovered between newly sequenced proteins and various classes of proteins in these databases. The sequence similarities detected, though often quite unexpected, have had important ramifications. Some examples include: (i) the homology (*1*) between bovine cyclic AMP (adenosine 3′,5′-monophosphate)–dependent kinase and the Rous avian and Moloney murine sarcoma virus src proteins (*2*). This finding reinforced the hypothesis that the *src* genes originated in host genomes and was the first evidence of similarity between an oncogene product and a cellular protein of known function. (ii) The near identity between the v-*sis* oncogene product and platelet-derived growth factor (*3*). The similarity was so strong that it is likely that the chromosomal *sis* gene codes for a growth factor. This serendipitous finding has stimulated new interest in the role of growth factors in oncogenesis. (iii) The similarity of the T-cell receptor protein to immunoglobulin proteins (*4*).

David J. Lipman is a Senior Staff Fellow of the Mathematical Research Branch of the National Institute of Arthritis, Diabetes, and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20205. William R. Pearson is an Assistant Professor in the Department of Biochemistry, University of Virginia, Charlottesville 22908.

In the past, identifying new proteins through database searches has been difficult. Computer programs required several hours on a minicomputer, or made important compromises in the sensitivity and selectivity of the search. One of the most rigorous programs for comparing amino acid sequences, SEQHP (*5*), requires more than 8 hours to compare a 200-residue protein to the 500,000-residue NBRF (National Biomedical Research Foundation) protein library on the VAX 11/750 computer.

We have developed an algorithm, used in the computer program FASTP, in which we can compare a 200-residue sequence to the NBRF library in 2 to 5 minutes on the VAX 11/750. The program searches rapidly because it first screens sequences for similarity by looking for aligned identical amino acids. It is sensitive because it considers conservative amino acid replacements as well as identities. FASTP has been modified to run on a wide variety of computers including the IBM PC. In this article, we discuss the basis of the algorithm and its application to two proteins evolutionarily related to other sequences in the database. In addition, we show an example of a search which presented puzzling results and discuss criteria for evaluating such results.

FASTP was written in the "C" programming language, originally on a VAX 11/780 with the UNIX operating system. It has since been moved to the VAX/VMS operating system, and to an IBM PC microcomputer. Memory requirements for the microcomputer are modest, but the program needs a disk drive with sufficient capacity to hold the protein sequence library (*6*).

The algorithm. Most DNA and protein sequence similarity algorithms compare each nucleotide or amino acid of one sequence with all of the residues in the second sequence. With these algorithms, comparison of a 200-amino-acid sequence to the 500,000-residue protein library requires approximately 10^8 comparisons. Wilbur and Lipman (*7*) described an algorithm that permits rapid searches of protein and nucleic acid databases by focusing only on groups of identities between the two sequences and therefore requires fewer comparisons in a search. We have modified that algorithm so as to improve its sensitivity and efficiency.

In the original algorithm, identities or groups of identities are rapidly located with a tool known in computer science as a lookup table (*8*). As an example, we use this technique to locate all the identities between two amino acid sequences (*9*):

Sequence number	Position						
	1	2	3	4	5	6	7
1	F	L	W	R	T	W	S
2	S	W	K	T	W	T	

First, a table of the positions is constructed for each type of amino acid (ktup = 1) or amino acid pair (ktup = 2) in sequence 1. The table would list L as appearing at residue 2, W at 3 and 6, R at 4, and so forth. Comparison proceeds by looking up each of the amino acids of sequence 2 in the table of amino acid positions of sequence 1. Thus, the S at residue 1 of the second sequence occurs at residue 7 of the first sequence, W of the second appears at 3 and 6 in the first, and so on. For a pair of sequences, the number of comparisons required is much closer to the sum of the lengths of the two sequences than the product.

In conjunction with the lookup table, Wilbur and Lipman (*7*) developed the "diagonal" method for locating regions

of similarity between two sequences which is based on identities and allows mismatches but not insertions or deletions. For each matched residue (or pair of residues) found with the lookup table, the difference between the position of the match in the two sequences (the offset) is calculated. The offsets will be equal for matches which may be simultaneously aligned without introducing gaps. A diagonal line in a dot-matrix homology plot (*10*) is composed of identities sharing the same offset.

In the above example, the S in position 1 of sequence 2 matches S in sequence 1 at an offset of 7 − 1 = 6; the W of position 2 matches at offsets 1 and 4; the T at 1; W of position 5 at 1 and −2, and T at −1. At an offset of 1, there are three matches, while each of the other offsets has one or zero match. The diagonal method compares the two sequences by scanning sequence 2 once from beginning to end; the score in an offset is increased for each identity and decreased for each mismatch. Those offsets with the highest scores represent local regions of similarity between the two sequences. At this stage, the Wilbur and Lipman algorithm (*7*) computes a final similarity score, allowing for insertions and deletions, based on all those identities or blocks of identity located within a preset distance from any of these local regions of similarity.

We have developed a new algorithm that uses a modified form of the diagonal method and dramatically improves the efficiency and sensitivity of amino acid sequence comparisons. The new algorithm locates the beginning and end positions in both sequences of the five regions of highest similarity found by the diagonal method. In the example, the best region of similarity is at an offset of 1 and extends from residue 3 to 6 of sequence 1. The new algorithm then takes advantage of the fact that amino acid replacements occur far more frequently than insertions or deletions (*11*). The five highest scoring local regions are rescored by comparing the similarity of all the paired amino acids, replacements as well as identities, using an amino acid replaceability matrix, the PAM250 matrix (*11*). Aligned identical amino acids which are rare (such as cysteine and tryptophan) receive higher scores than identities among more common amino acids (such as serine and alanine), and replacements which have occurred frequently in evolution, such as methionine → leucine, also receive positive scores while unlikely substitutions (such as cysteine → tryptophan) receive negative scores. In the example, the aligned W and T residues get scores of 17 and 3, respectively, and the K → R substitution gets a score of 3, for a total score of 40. By not considering insertions or deletions at this stage, computation time and memory requirements are greatly reduced while sensitivity is significantly improved.

The score from the best sub-sequence alignment, based on the PAM250 matrix, is used as the similarity score between the two sequences (the initial score). When performing a database search, an initial score is calculated between the test sequence (the query sequence) and

Fig. 1. Comparison of bovine cyclic AMP–dependent kinase with the NBRF protein sequence library. FASTP was used to compare the protein sequence of bovine cyclic AMP–dependent kinase, OKBO2C, with each of the sequences in the National Biomedical Research Foundation protein sequence library. The ktup parameter was set to 2 for this search. (A) A histogram indicating the number of database sequences (second column of numbers) at any given initial score (first column). Thus, there was one database sequence whose initial score was less than 2, while there were four sequences whose scores were between 6 and 8. The horizontal double bars indicate the approximate number of sequences at each initial score. The number in parentheses after "mean score" is the standard deviation. (B) The alignment of OKBO2C(′) with two sequences from the library search(″). The boundaries of the initial alignment found during the library scan are marked with vertical lines. The optimized alignment is denoted by a colon for an identity and a dot for a conservative replacement. Insertions made during optimization are marked with a dash.

each of the sequences in the database. The initial scores are displayed in a histogram and a mean and standard deviation are calculated. The sequences are then ranked by their initial score. In addition, an optimized score, which allows for insertions and deletions, is also computed for the high ranking similarities. The optimized alignment uses a modification of the method first described by Needleman and Wunsch (*12*) with subsequent refinements (*5, 13*). The Needleman and Wunsch optimization methods belong to a class of dynamic programming algorithms which have been used in speech processing, computer science, text collation, and the analysis of bird song (*14*). Our algorithm performs a local optimization, so that dissimilar portions of the sequence outside the optimized alignment region do not affect the score of the alignment (*15*).

FASTP may be operated in either of two modes. The algorithm may compute the initial score using dipeptide matches (ktup = 2) or individual identities (ktup = 1). The search is considerably faster with ktup = 2 because, in using the lookup table, only pairs of identities are considered. These occur rarely between randomly related proteins. Greater sensitivity is generally obtained however with ktup = 1 (*16*).

Example 1: Bovine cyclic AMP–dependent kinase. In this example (Fig. 1), bovine cyclic AMP–dependent kinase, OKBO2C, was compared to the 2677 protein sequences in the NBRF protein sequence library, with ktup = 2 (*17*). The histogram of initial scores (Fig. 1A) illustrates that there were very few protein sequences with scores between 50 and 75 (5 to 9 standard deviations above the mean), but there were 15 scores higher than 75 (including the self-comparison). Table 1 contains the initial and optimized scores for the 20 most similar sequences. Note that the optimized scores of the 16 proteins with the highest initial scores often double when gaps are allowed, but that the initial scores for three unrelated proteins increase very little, if at all.

Optimized alignments of OKBO2C with a related *src* gene sequence (TVFV-R) and with an unrelated fetal globin chain are also shown (Fig. 1B). In each alignment, the ends of the region which determined the initial score are marked by vertical lines. Because FASTP uses an algorithm to find the best local similarity, often the optimized alignment region does not extend from the beginning to the end of either protein sequence. Thus in the comparison between bovine kinase and fetal hemoglobin, the optimized alignment is within the boundaries of the initial alignment. Other aligned identical amino acids were not marked because the dissimilarity of the intervening amino acid pairs prevented the alignment from being extended. FASTP clearly distinguished all the tyrosine kinase–related sequences from the rest of the database.

Example 2: Rat angiotensinogen pre-

Table 1. Protein sequences similar to bovine cyclic AMP–dependent protein kinase.

Library sequence matched*		Score†	
Identifier	Protein	I	O
OKBO2C	Bovine cyclic AMP–dependent protein kinase	1810	1810
TVYUH	Avian erythroblastosis virus kinase-related transforming protein (erbB)	96	179
GNMVRR	Feline sarcoma virus gag-fgr polyprotein	90	191
TVBY8	Yeast cell division control protein 28	88	224
TVFV-R	Rous sarcoma virus kinase-related transforming protein (src)	86	217
TVCHS	Chicken kinase-related proteins 1 and 2 (src)	86	212
TVFV60	Rous sarcoma virus (avian sarcoma virus) kinase-related transforming protein (src)	86	216
GNFVG9	Avian sarcoma virus gag-yes polyprotein p90	86	203
TVRT M	Rat kinase-related transforming protein (mos)	85	112
GNVWGM	Abelson murine leukemia virus gag-abl polyprotein	83	237
GNFVF	Fujinami sarcoma virus gag-fps polyprotein	80	190
TVHU-T	Human probable kinase-related protein (mos)	78	133
TVMV M	Moloney murine sarcoma virus kinase-related transforming protein (mos)	76	111
TVMV1M	Moloney murine sarcoma virus (strain m1) kinase-related transforming protein (mos)	76	110
TVMS M	Mouse kinase-related protein (mos)	76	108
GNMVCS	Feline sarcoma virus gag-fes polyprotein	62	161
GNMVGC	Feline sarcoma virus gag-fes polyprotein	62	163
QQECUC	*E. coli* unc hypothetical protein C-130	51	51
CYBOB	Bovine beta crystallin Bp chain	51	54
HBGTF	Goat and sheep hemoglobin beta fetal chain	49	49

*Identifiers and protein descriptive names are those used by the National Biomedical Research Foundation, Washington, D.C.. †I, initial; O, optimized.

Table 2. Protein sequences similar to rat angiotensinogen measured with ktup = 1 and ktup = 2.

Library sequence matched		Score*			
		ktup = 1		ktup = 2†	
Identifier	Protein	I	O	I	O
ANRT	Angiotensinogen precursor, rat	2294	2294	2294	2294
ITHU	Alpha-1-antitrypsin precursor, human	96	288	(32	288)
ITBA	Alpha-1-antitrypsin, baboon	95	285	(34	285)
ANHU	Angiotensinogen, human (fragment)	78	80	78	80
ANHO	Angiotensinogen, horse, bovine and chicken	78	78	78	78
G1HUNM	Ig heavy chain V-2 region, human	74	75		
DXCH	Gene X protein, chicken (fragment)	65	202	65	202
XHHU3	Antithrombin-III precursor, human	64	229	52	229
DYCH	Gene Y protein (ovalbumin-related), chicken	63	235	63	235
ITHUC	Alpha-1-antichymotrypsin precursor, human	62	234	(25	30)
TVMV-S	PDGF-related transforming protein (sis)	59	68		
HAMS	Hemoglobin alpha chains, mouse	56	63		
OACH	Ovalbumin, chicken	58	222	58	222
ODAS1	Cytochrome oxidase polypeptide	56	57		
K1HUNY	Ig kappa chain V-1 region, human	56	56		
HATG2	Hemoglobin alpha-2 chain, echidna	56	60		
TSBYAB	Tryptophan synthetase, baker's yeast	56	59		
G3HUWI	Ig gamma-3 heavy chain disease proteins	56	59		
TSECB	Tryptophan synthetase beta chain	55	71		
TSEBBT	Tryptophan synthetase beta chain	55	72		

*I, initial; O, optimized. †Values in parentheses were not ranked in the 20 highest scores with ktup = 2. They are included for comparison with the ktup = 1 results.

cursor. In the first example, oncogene proteins similar to bovine cyclic AMP–dependent kinase were clearly resolved from the bulk of the protein library. The similarity between the kinase and the oncogenes was strong, and there was little difference in the results when using ktup = 2 as compared with ktup = 1. This is usually, but not always, the case. The next example shows an important difference in the results between a ktup = 1 as compared to a ktup = 2 search. Here the query sequence is rat angiotensinogen (ANRT), which is related to the ovalbumin, antitrypsin, and antithrombin III family (*18*). Nine of the ten highest scoring sequences were members of the expected family (gene X and gene Y proteins are closely related evolutionarily to ovalbumin), and ovalbumin was ranked slightly lower (Table 2). The optimization clearly distinguished ovalbumin from its closely ranked neighbors.

The results of a ktup = 2 search were quite similar with one important exception: the antitrypsin proteins were not among the top ranked sequences. The explanation for this is found in Fig. 2, which displays the optimized alignment of the rat angiotensinogen precursor and human α-1-antichymotrypsin precursor (ktup = 1). Although a number of identities were seen throughout the alignment, they were dispersed fairly evenly and very few appeared in pairs. In these situations, the initial stage of the algorithm with ktup = 2 would miss potentially significant relationships. Different situations can be found and investigators will frequently want to compare the results of searches with ktup = 2 and ktup = 1.

Table 3. Statistical significance (z value) of protein similarity scores. Protein sequences from the searches discussed in examples 1, 2, and 3 were compared with the best related and unrelated library sequences found. *Z* values [(score − mean score)/standard deviation] were calculated for the initial score from the mean and standard deviation of the database initial scores (initial scan), and for the initial (I) and optimized (O) scores from the mean and standard deviation of scores against randomly permuted versions of the database sequence in question. In the latter case, 50 comparisons (ktup = 1) were made with shuffled sequences.

Query sequence	Library sequence matched		Initial scan	Randomized	
	Identifier	Protein		I	O
OKBO2C (bovine cyclic AMP kinase)	TVBY8	Yeast cell cycle control	11.3	11.1	24.9
	TVFV-R	Src	11.0	10.4	23.6
	TVMS M	Mos	9.3	7.6	10.1
ANRT (rat angiotensinogen)	ITHU	Alpha-1 antitrypsin	10.0	8.3	25.8
	G1HUNM	Human Ig heavy chain (V-2)	6.8	8.1	7.8
	XHHU3	Antithrombin	5.3	5.1	24.1
	ITHUC	Alpha-1 antichymotrypsin	5.0	3.9	15.5
	TVMV-S	PDGF-related sis	4.6	3.8	3.8
VHVUNH (snowshoe hare bunyavirus nucleoprotein)	ORBPL	Lambda replication protein	4.4	4.0	2.5
	GHRB	Rabbit Ig gamma C region	4.3	4.8	3.6

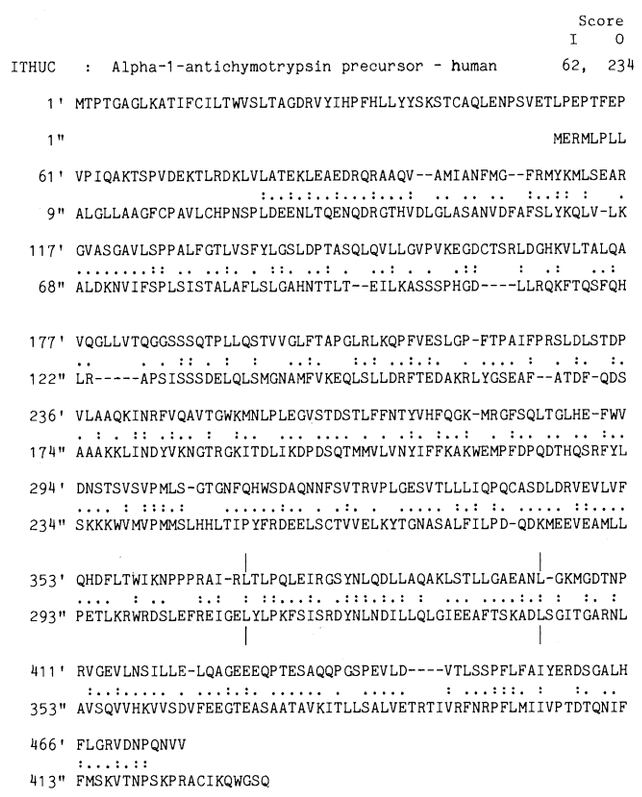

Fig. 2. Optimized alignment of rat angiotensinogen (ANRT) with human alpha-1 - antichymotrypsin precursor (ITHUC), ktup = 1.

Example 3: Probable nucleoprotein, snowshoe hare bunyavirus. Many newly sequenced proteins will be unrelated to sequences currently available. One such example is the probable nucleoprotein of snowshoe hare bunyavirus (VHVUNH). The histogram generated from this search using ktup = 1 and the top five scores are shown (Fig. 3, A and B). We shall focus on the potential relationship between the nucleoprotein and the immunoglobulin (GHRB). Although the initial score for GHRB was among the highest in the database, it was not as far from the other scores as were the similarities discussed in the previous examples. Furthermore, as with the other highly ranked sequences, there was little change in score with optimization. The alignment with GHRB paired the carboxyl terminus of the nucleoprotein with the amino terminus of the immunoglobulin (Fig. 3C), which is an unusual pattern of conservation between homologous proteins. Although the immunoglobulins were heavily represented in the database, no other immunoglobulins appeared among the high-scoring sequences. In contrast to this, when the T-cell receptor was searched against the database, all of the high-scoring sequences were from the immunoglobulin family. It would seem likely, therefore, that this database search did not detect any sequences with a biologically significant similarity to the bunyavirus nucleoprotein.

Evaluation of statistical significance. The significance of a relationship found with a database search depends on the relative magnitude of the initial score, the change in score after optimization, and the alignment itself. Although one of the great strengths of this search tool is its ability to find unexpected relationships, it is obvious that one must apply any available biological information in evaluating a potential relationship. However, when the biological context of a

potential relationship is puzzling, or where there are obvious biological reasons for a similarity yet the other criteria are less convincing, an estimate of statistical significance can be useful.

At present, the most satisfactory method of obtaining an estimate of statistical significance is to compare the query sequence with randomly permuted versions of the potentially related sequence. Each of the generated sequences retains the exact length and amino acid composition of the original database sequence. We have written a second program, RDF, to perform this comparison (*6*). RDF determines both the initial and optimized scores in each comparison so that the statistical significance of both the initial score and the optimized score can be evaluated.

The similarity scores determined by our algorithm, and by other sensitive methods, do not follow a normal distribution from which one can calculate a probability (*P* value). As a result, the significance of a similarity is frequently expressed not as a *P* value but as a *z* value (*7, 11, 19*) where

$$z = (\text{similarity score} - \text{mean of random scores})/(\text{standard deviation of random scores})$$

Table 3 summarizes the *z* values for a number of similarity scores between related and unrelated proteins. As expected, the *z* values from the initial scores of randomly shuffled sequences were very close to *z* values calculated with the mean and standard deviation of initial scores from the database scan. This may not be the case if the query sequence has a highly nonuniform amino acid composition or if the matching library sequence is extremely long. For those relationships believed to be biologically significant, the *z* value for the initial score was greater than 4 and that for the optimized score was even higher. From our experience with a large number of database searches we use the following guidelines:

z value of initial and optimized score	
$z > 3$	possibly significant
$z > 6$	probably significant
$z > 10$	significant

These values are similar to those suggested by other investigators (*20*). Empirical and theoretical work is under way to develop more definitive estimates of the statistical significance of sequence similarities (*21*).

Sensitivity and selectivity. Ideally, a protein similarity search algorithm should be highly selective and highly sensitive. Such a method would consistently rank all database sequences with a statistically significant similarity to the query sequence above any other sequences. Efforts to increase sensitivity, such as allowing for gaps or conservative amino acid substitutions carry with them the risk of decreasing selectivity by increasing the scores of unrelated sequences (*22*). Furthermore, for the optimal balance of sensitivity and selectivity, the best value for gap penalty or the best values for a replaceability matrix may vary with different query sequences.

Perhaps the best test of a search method is how well it identifies the known relatives of a given protein. In the classification scheme devised by Dayhoff (*11*) and employed in the NBRF database, the largest grouping of proteins is the superfamily. Proteins grouped in a superfamily generally have statistically significant sequence similarity and thus presumably share a common evolutionary origin. Within a superfamily, proteins may be clustered into families if they share approximately 50 percent sequence similarity. Proteins within a superfamily may have acquired quite different functions but often share similar three-dimensional structure. Proteins within a family generally share similarity in function as well.

We used FASTP to compare the human cytochrome c sequence against a subset of the protein sequence library containing all of the members of the cytochrome c superfamily. All of the eukaryotic cytochrome c sequences had initial scores higher than 160. The lowest was a *Tetrahymena* cytochrome c with

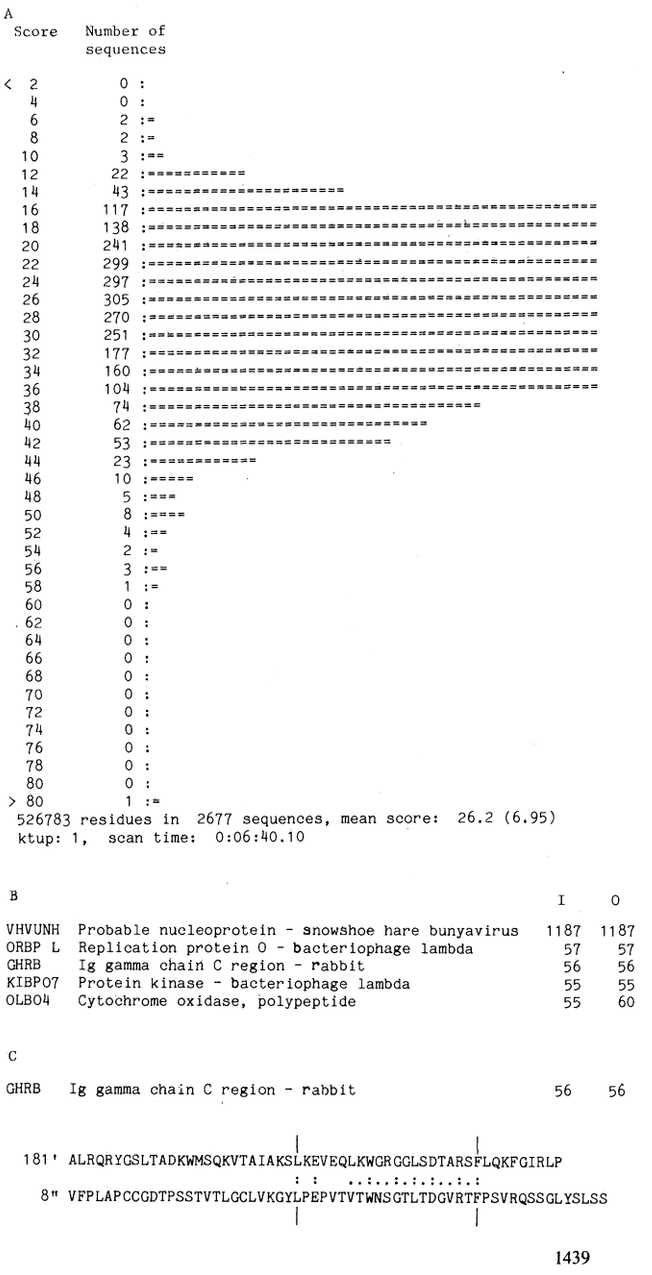

Fig. 3. A protein sequence with no significant similarities—snowshoe hare bunyavirus nucleoprotein. The protein sequence library was scanned using snowshoe hare bunyavirus nucleoprotein as the query sequence with ktup set to 1. (A) The histogram of initial scores. (B) The best five matching sequences. (C) Alignment with rabbit immunoglobulin gamma chain C region (GHRB).

an initial score of 165, optimized to 230. Initial scores for the 29 prokaryotic cytochrome c's ranged from 14 to 291; 11 prokaryotic sequences had scores higher than 60.

Cytochromes show little change in length and it is perhaps not surprising that very distant homologies can be recognized with our method. The serine protease family of enzymes has evolved at more than twice the rate of the cytochrome c family (*23*), and the enzymes show more changes in chain length. When bovine trypsinogen, TRBOTR, was compared to the 35 members of the serine protease superfamily, 30 of the sequences had initial scores greater than or equal to 64, the lowest of which, human haptoglobin, increased to 175 with optimization (Table 4). The program clearly detected the relationship with *Streptococcus griseus* trypsin. Two of the lowest ranked sequences, *S. griseus* proteases A and B, are examples of enzymes with conserved serine protease active sites which have been identified by crystallography, not by protein sequence homology. The algorithm's sensitivity is evident in both examples. However, because statistically significant similarity may not be demonstrable between all pairs of proteins in a superfamily (*19*), it is necessary to be cautious in interpreting negative results in a database search.

A protein searching strategy. When possible, similarity searches with amino acid sequences are much more effective than nucleotide sequence searches. Several different nucleotide sequences can encode a single amino acid. Furthermore, because there is data concerning amino acid replacement frequencies in evolution, similarities in amino acid sequence may be detected even in the presence of complete dissimilarity in nucleotide sequence. A generally applicable strategy for comparing a query sequence with the protein database might be: (i) search the database with ktup = 2. Genuinely related sequences will have initial scores greater than 50 that increase to between 100 and 300 after optimization. The z value of the optimized score will be 10 or greater. If there are no clearly related sequences, then (ii) search again with ktup = 1. At least the 20 sequences with highest scores should be evaluated in terms of the magnitude of the initial score, change of score with optimization, the quality of the alignment, and the biological context of a potential relationship with the query sequence.

At this point, in most instances, the analysis is complete. Either a protein with a biological relationship to the query has been found or the search is negative. In some cases the picture is less clear, and an evaluation of statistical significance is required. If the resulting z values are greater than 6 or less than 3, interpretation of the results is usually straightforward. There will be a number of relationships found whose z values are between 3 and 6 which may be interesting to an investigator. Our experience with a large number of database searches has made us conservative; we consider most such relationships to be speculative. There are more than 2600 protein sequences in the NBRF database, and scores with z values greater than 3 are common. Table 3 includes a comparison of rat angiotensinogen and the variable region of a human immunoglobulin heavy chain, with z values around 8, even though a biologically significant relationship between the two proteins is highly unlikely.

As our understanding of protein evolution and the relationship between sequence and structure grows, new methods of comparison will be needed. While we have found FASTP to be of great utility in a wide variety of cases, replaceability matrices constructed for specific protein families might be useful, or different matrices might be used for specific regions of a protein (such as the active site). These improvements should further enhance the ability to extend our knowledge about newly sequenced proteins by means of information stored in protein sequence databases.

Table 4. Comparison of bovine trypsinogen with members of the serine protease superfamily. Thirty-five sequences from the serine protease superfamily were compared with bovine trypsinogen. Some high-scoring duplicated sequences are not shown.

Identifier	Protein	Score*	
		I	O
TRBOTR	Trypsinogen, bovine	1112	1112
TRPGTR	Trypsinogen, pig	998	998
TRDF S	Trypsinogen, spiny dogfish	455	846
NGMSG	7*S* Nerve growth factor gamma chain	310	537
KQMS1	Prokallikrein mGK-1 precursor, mouse	306	519
KQMSM	Kallikrein, submaxillary gland, mouse	273	341
TRCY1	Trypsin I, crayfish	253	446
PRRTG	Group-specific protease, rat	204	315
KCUF	Collagenolytic protease, fiddler crab	181	379
KQPG	Kallikrein, pig (fragments)	181	259
KFHU	Factor IX precursor, human	144	395
KXBO	Protein C, human	143	363
UKHUT	Plasminogen activator tissue-type, human	137	217
PLHU	Plasminogen, human	134	410
EXBO	Factor X (Stuart factor), bovine	128	335
KYBO A	Chymotrypsinogen A, bovine	122	434
ELPG	Elastase, pig	102	327
TBHU	Prothrombin, human	99	180
TRSM G	Trypsin, *Streptomyces griseus*	91	263
HPHU2	Haptoglobin-2 precursor, human (fragment)	64	179
HPHU1	Haptoglobin-1, human	64	175
BBHUB	Complement factor B, Bb fragment, human	50	84
PRSMBG	Protease B, *S. griseus*	41	42
PRSMAG	Protease A, *S. griseus*	37	37
TRYXB4	Alpha lytic protease, *Myxobacter*	33	39
BZSO	Streptokinase, *Streptococcus*	20	28

*I, initial; O, optimized.

References and Notes

1. In this article, homology means that two proteins share a common ancestral protein.
2. W. Barker and M. Dayhoff, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2836 (1982).
3. R. F. Doolittle *et al.*, *Science* **221**, 275 (1983); M. D. Waterfield *et al.*, *Nature (London)* **304**, 35 (1984).
4. S. Hedrick *et al.*, *Nature (London)* **308**, 153 (1984); Y. Yanegi *et al.*, *ibid.*, p. 145.
5. W. Goad and M. Kanehisa, *Nucleic Acids Res.* **10**, 247 (1982).
6. Copies of FASTP and RDF (for evaluating statistical significance) with complete source code are available from D.J.C. and W.R.P.
7. W. J. Wilbur and D. Lipman, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 726 (1983); *SIAM (Soc. Ind. Appl. Math.) J. Appl. Math.* **44**, 557 (1984).
8. The first application of a lookup table to biological sequences was reported by J. P. Dumas and J. Ninio [*Nucleic Acids Res.* **10**, 197 (1982)].
9. The following abbreviations were used for amino acids: A, alanine; C, cysteine, D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; and Y, tyrosine.
10. J. Maizel and R. Lenk, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 7665 (1981).
11. M. Dayhoff, *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Silver Spring, Md., 1978), vol. 5, supplement 3. The PAM250 matrix was derived by comparing various closely related proteins and determining the probability that a given amino acid would be replaced by any of the other amino acids over a fixed period of time. Empirical studies have shown that this matrix allows the detection of more distant relationships than if only identities or minimum base changes are considered.
12. S. Needleman and C. Wunsch, *J. Mol. Biol.* **48**, 443 (1970).
13. P. Sellers, *SIAM (Soc. Ind. Appl. Math.) J.*

Appl. Math. **26**, 787 (1974); D. Sankoff, *Proc. Natl. Acad. Sci. U.S.A.* **69**, 4 (1972); T. Smith and M. Waterman, *J. Mol. Biol.* **147**, 195 (1981).
14. J. Kruskal, *SIAM (Soc. Ind. Appl. Math.) Rev.* **25**, 201 (1983).
15. The algorithm used to calculate the optimized alignment is considerably faster than the normal Needleman-Wunsch procedure, because it only compares the two sequences along a narrow band centered on the initial alignment. With this restriction, gaps may be inserted in either sequence so as to include areas of similarity up to 16 residues from the offset of the initial alignment. Some proteins will not be completely aligned with this limited local optimization because the method focuses on the most similar region; nevertheless, scores of related proteins increase substantially.
16. Search time is approximately linearly related to the length of the query sequence. A search with ktup = 1 generally takes four to five times as long as a ktup = 2 search. Search times with ktup = 2 on a VAX 11/750 computer range from 1.5 minutes for cytochrome c (104 amino acids) to 2.9 minutes for angiotensinogen (477 amino acids) to 4.3 minutes for mouse epidermal growth factor precursor (1247 amino acids). When ktup = 1, the search times increase to 3.8, 12.8, and 26.8 minutes, respectively. Searches take roughly four times longer on an IBM PC microcomputer.
17. All amino acid sequences used in this work were extracted from the NBRF database, and the protein identifiers, such as OKBO2C, are those used by this database.
18. L. Hunt and M. Dayhoff, *Biochem Biophys. Res. Commun.* **95**, 864 (1980); R. Doolittle, *Science* **222**, 417 (1983).
19. R. Doolittle, *Science* **214**, 149 (1981).
20. R. Doolittle (*19*) uses a value of $z > 3$ for significance by means of a more selective measure based solely on aligned identities, while investigators at the Protein Identification Resource, NBRF, typically use a value of $z > 5$ with the PAM250 replaceability matrix (W. Barker, personal communication). We suggest higher z values because of the local nature of our scoring method.
21. D. Sankoff and R. Cedergren, *J. Mol. Biol.* **77**, 159 (1973); J. Steele, *SIAM (Soc. Ind. Appl. Math.) J. Appl. Math.* **42**, 731 (1982); S. Karlin, G. Ghandour, F. Ost, S. Tavare, L. Korn, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 5660 (1983); D. Lipman, W. Wilbur, T. Smith, M. Waterman, *Nucleic Acids Res.* **12**, 215 (1984).
22. We have compared FASTP to SEQHP (*5*). SEQHP is a very sensitive program which uses a derivative of the Needleman-Wunsch algorithm and the PAM250 scoring matrix to determine local protein sequence similarities. Because of the computational requirements of the Needleman-Wunsch algorithm, this program requires almost two orders of magnitude more time for a database search. We have found SEQHP to be slightly more sensitive and significantly less selective than FASTP. In a typical example, SEQHP found four sequences with scores between 86 and 91, while FASTP found only two of these four. However, the z values for the two sequences not found by FASTP were 3.5 and 3.7, while the two scores found by FASTP (actually ranked lower by SEQHP) had z values of 18 and 10. The high scores were due to long runs of conservative replacements with much less statistical significance.
23. A. Wilson, S. Carlson, T. White, *Annu. Rev. Biochem.* **46**, 573 (1977).
24. We thank W. Barker of the National Biomedical Research Foundation for help with the protein sequence library and A. Furano for his suggestions on the manuscript. We would also like to thank the Department of Applied Mathematics and Computer Science at the University of Virginia, and A. Batson, for the generous gift of computer time, and the Division of Computer Resources and Technology, NIH, for the use of an IBM PC microcomputer. Supported by biomedical research support grant SO7-RR05431 to the University of Virginia School of Medicine and by the National Institute of Arthritis, Diabetes, and Digestive and Kidney Diseases, NIH.

14 September 1984; accepted 30 November 1984

AAAS—Newcomb Cleveland Prize
To Be Awarded for an Article or a Report Published in *Science*

The AAAS-Newcomb Cleveland Prize is awarded annually to the author of an outstanding paper published in *Science*. The 1985 competition starts with the 4 January 1985 issue of *Science* and ends with the issue of 27 December 1985. The value of the prize is $5000; the winner also receives a bronze medal.

Reports and Articles that include original research data, theories, or syntheses and are fundamental contributions to basic knowledge or technical achievements of far-reaching consequence are eligible for consideration for the prize. The paper must be a first-time publication of the author's own work. Reference to pertinent earlier work by the author may be included to give perspective.

Throughout the year, readers are invited to nominate papers appearing in the Reports or Articles sections. Nominations must be typed, and the following information provided: the title of the paper, issue in which it was published, author's name, and a brief statement of justification for nomination. Nominations should be submitted to the AAAS-Newcomb Cleveland Prize, AAAS, 1515 Massachusetts Avenue, NW, Washington, D.C. 20005. Final selection will rest with a panel of distinguished scientists appointed by the editor of *Science*.

The award will be presented at a session of the AAAS annual meeting. In cases of multiple authorship, the prize will be divided equally between or among the authors.