Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 5509–5513, July 1990
Evolution

Protein database searches for multiple alignments

(homology/sequence comparison/statistical significance/alignment algorithms/pattern recognition)

STEPHEN F. ALTSCHUL AND DAVID J. LIPMAN

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

Communicated by Samuel Karlin, April 30, 1990

ABSTRACT **Protein database searches frequently can reveal biologically significant sequence relationships useful in understanding structure and function. Weak but meaningful sequence patterns can be obscured, however, by other similarities due only to chance. By searching a database for multiple as opposed to pairwise alignments, distant relationships are much more easily distinguished from background noise. Recent statistical results permit the power of this approach to be analyzed. Given a typical query sequence, an algorithm described here permits the current protein database to be searched for three-sequence alignments in less than 4 min. Such searches have revealed a variety of subtle relationships that pairwise search methods would be unable to detect.**

Protein and nucleic acid sequence comparison has become an important tool for molecular biologists. Frequently the first clues to the structure or function of a newly sequenced protein come from similarities it exhibits to other proteins that have already been studied. Protein database searches can draw the researcher's attention to unsuspected relationships and suggest future lines for investigation (1, 2). One problem is that while close relationships are easy to discover with existing tools, weaker ones may be indistinguishable from chance similarities that would be found between even random sequences. As the size of the protein database increases, the "noise" from chance similarities found in database searches grows as well, making weak relationships ever harder to detect.

One strategy for distinguishing a chance sequence pattern from a biologically relevant one is to search the database for several instances of it among otherwise dissimilar proteins. The basic principle is that while an unusual event may be due to chance, recurrence of the same event requires explanation: lightning striking twice in the same place probably indicates a lightning rod. In this paper, we discuss how recent statistical results (3, 4) allow us to clothe this intuition in numbers, so that a multiple alignment is seen to be statistically significant while none of the pairwise alignments that comprise it are. We also describe a simple computational strategy that allows a database to be searched for multiple alignments in reasonable time.

This approach is distinct from "profile analysis" (5) and related methods (6–8), which search for pairwise similarities to a sequence "profile" derived from a predefined multiple alignment. In contrast, our method takes a *single* query sequence and searches a standard database of individual protein sequences for segments with which it can form statistically significant multiple alignments. An obvious generalization would be to use a profile as the query. Using our method to search the National Biomedical Research Foundation Protein Identification Resource (PIR) protein database with a variety of query sequences, we have found many relationships that could not have been distinguished from chance by pairwise database search methods.

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

Measuring Local Similarity

To find patterns among protein sequences, it is useful to have some measure of sequence similarity. The usual approach is to assign scores to aligned pairs of amino acids and, if gaps are to be allowed, to amino acids aligned with nulls (missing residues). The similarity of a particular alignment is then the sum of these scores. Such similarity measures may be described as global or local, depending upon whether every residue of a sequence is required to participate (9, 10), or whether an alignment may be confined to segments of the proteins compared (11, 12). Generally, local measures such as the ones we will discuss are preferred for database searches, where cDNAs may be compared with partially sequenced genes, and where distantly related proteins may share only isolated regions of similarity (e.g., in the vicinity of an active site). In this paper we will consider only local alignments that lack gaps; the mathematical, conceptual, and algorithmic tractability that this buys will be found to repay the sacrifice in alignment generality.

We define a *segment* of a protein sequence to be a set of contiguous residues. A *subalignment* (or for simplicity *alignment*) of two sequences can be specified by choosing an equal-length segment from each; the alignment's score is then determined, as described above, by a matrix of amino acid *substitution scores*. The most widely used substitution scores are variations of the PAM matrix (13); we will use such scores in the applications that follow. An *optimal alignment* is simply the subalignment, of any length, with the highest possible score.

While it is obvious how to extend the above definition of alignment to three or more sequences, the score of such a multiple alignment may be defined in a variety of ways (14). For simplicity, we will take the score of a multiple alignment to be the sum of the scores of all the pairwise alignments it imposes; we call these SP scores, for "sum of the pairs." The discussion that follows applies as well to many other possible choices for multiple alignment scores.

Statistical Significance of Pairwise and Multiple Sequence Alignments

To decide whether a given alignment can reasonably be explained by chance, one needs to have a model of chance. A protein can be modeled most simply as a random sequence of independently chosen amino acids, with the different types of residue occurring with certain underlying frequencies. Given two such "random protein sequences," what is the probability P that the optimal alignment score (as defined above) will be at least S? Recent mathematical results answer this question (3, 4). The theory involves two parameters, λ and K, for which explicit formulas are given, that are depen-

Abbreviation: PIR, Protein Identification Resource.

dent upon the scoring scheme and the underlying amino acid frequencies. The probability P is then well approximated by the formula

$$1 - \exp(-KNe^{-\lambda S}), \quad [1]$$

where N is the product of the lengths of the two sequences. For the comparison of three random protein sequences, the chance that a three-way alignment (3-alignment) will have score at least S is given by the same formula, except that different parameters λ and K must be used, and N is the product of the three sequences' lengths. The results generalize to an arbitrary number of sequences.

There are several limits to the applicability of formula 1; one is that the lengths of the sequences compared should not be too dissimilar. In this instance, however, the formula overestimates P, yielding a conservative estimate of statistical significance. The formula is useful primarily for providing a rough benchmark for assessing the significance of sequence alignments, since the random protein model is only an approximation.

Formula 1 can be used to get some idea of the power to be gained from searching a database for multiple as opposed to pairwise alignments. For pairwise alignments, the N of the formula is just the product of the length m of the query sequence and the length n of the database, i.e., the number of residues it contains. (Because of the redundancy found in many databases, taking n at a fraction—about one-half—of its true value often yields a better estimate of statistical significance.) When a database is searched for *two* segments to align with a segment of the query, $N = mn^2/2$. One divides by 2 because the segments chosen from the database may be ordered in two ways. For a 4-alignment, where three segments are chosen from the database, $N = mn^3/6$. If one assumes a specific random protein model and SP scores based on a PAM-120 version of the Dayhoff PAM matrix (13), Table 1 shows the parameters λ and K derived for two-, three-, four-, five-, and six-sequence comparison (3, 4). Also shown are the smallest scores significant at the 95% level when a 4,000,000-residue protein database is searched with a 150-residue query. Since each multiple alignment score is the sum of the imposed pairwise scores, we can compare the average pairwise score required to achieve significance. As expected, this number decreases as the number of segments in the alignment increases. Thus, for example, the database search described is expected to turn up about 15 pairwise alignments with score at least 47 purely by chance. (The expected number of distinct random alignments with score at least S is well approximated by the formula $KNe^{-\lambda S}$.) It is surprising, however, to find two segments that each can pair with the same query segment to yield a score of 47, and likewise with one another. This is the rationale for seeking multiple alignments.

Searching for Multiple Alignments

Given a query sequence, suppose we have somehow managed to find a 3-alignment (involving it and two sequences from a database) with a highly significant score. This does not necessarily constitute evidence that all the segments in the alignment are related. The high score may be due primarily to the similarity of just two related segments, while the third may bear only a chance resemblance to both. One way to gauge whether this is the case is to compare the P value (calculated as described above) of each implicit pairwise alignment with the P value of the 3-alignment. Only if the 3-alignment appears more significant than all the pairwise alignments it includes does the alignment provide real evidence for a three-sequence relationship. In practice this means that for discovering new relationships, 3-alignments are useful only when they comprise segments that are roughly equidistant; this provides the basis of the following algorithm. For purposes of discussion, we will assume the algorithm is applied to the database search described in Table 1; for different-sized searches, the numbers change but the principles do not.

Table 1. Statistically significant multiple alignment scores for searching a 4,000,000-residue protein database with a 150-residue query

Segments in alignment	Parameters		95% significance level	Average score per segment pair
	λ	K		
2	0.336	0.179	64	64
3	0.255	0.177	141	47
4	0.202	0.137	246	41
5	0.165	0.096	384	38
6	0.137	0.068	577	37

A pairwise alignment with score over 63 can be considered significant in its own right. A pairwise alignment with score much less than 47 is unlikely to participate usefully (as described above) in a significant 3-alignment. It is on pairwise alignments with scores between these bounds (from, say, 35 to 63) that we will concentrate. These alignments have sometimes been said to inhabit a "twilight zone": perhaps meaningful, but not clearly distinguishable from chance. Theory predicts that about 850 such alignments should be found in a search of a "random" database. Locating these alignments is the first step of our strategy. A brute-force algorithm for this purpose requires time proportional to the product of the lengths of the query sequence and the database. Running on a Sun 4, this requires approximately 1 μsec per residue squared, which translates into about 10 min for the search described. Parallel-architecture computers or special-purpose chips can reduce the search time by 2 orders of magnitude (15). Alternatively, fast heuristic search strategies can achieve over an order of magnitude greater speed on standard machines, but at the price of missing an occasional low-scoring alignment. For example, an implementation of the BLAST algorithm (2) can search a database at about 20 times the rate of a brute-force algorithm, while missing only about 3% of alignments with score 40, 1% of those with score 47, and an even smaller percentage of higher-scoring alignments.

Each of the saved pairwise alignments implies a specific *phase* for aligning the query and a database sequence. For example, residue 1 of the query may align with residue 8 of the database sequence, residue 2 with residue 9, etc. Two different pairwise alignments imply a phase for aligning the query with two different database sequences. These three sequences, locked into a specific phase, can be searched straightforwardly for a high-scoring 3-alignment in time linear in the length of the query sequence. The approximately 850 saved pairwise alignments imply approximately 350,000 such three-sequence sets. The second step of our strategy involves searching all these sets for high-scoring 3-alignments; the time is proportional to the square of the number of pairwise alignments saved. The example we have been discussing requires approximately 2 min on a Sun 4. Therefore, using the BLAST search strategy (2) for the first step, the complete algorithm can be executed in under 4 min on a Sun 4 for typical-length protein sequences and the current PIR database.

Preprocessing can reduce the time for the second step of the algorithm to about half a minute. By comparing the entire database against itself, and recording for every pair of sequences the highest pairwise similarity found, most pairs of alignments examined in the second step can be discarded as uninteresting after a single lookup. Usually this reduces the time for the complete algorithm to under 2 min.

Table 2. High-scoring pairwise alignments from a PIR database search with human α_1B-glycoprotein (OMHU1B)

PIR code	Protein description	Score	*P* value
S00474	Kinase-related transforming protein (kit) precursor - Mouse	58	0.52
K3HUVH	Ig kappa chain precursor V-III region - Human Vh	57	0.65
WGSMHH	Hygromycin B phosphotransferase - Streptomyces hygroscopicus	56	0.77
A25399	Antennapedia homeotic protein - Fruit fly	56	0.77
A25400	Antennapedia homeotic protein - Fruit fly	56	0.77
A23450	Antennapedia homeotic protein - Fruit fly	56	0.77
GNWVY	Genome polyprotein - Yellow fever virus (strain 17D)	54	0.95
XFSMF	Plasminostreptin (PSTI-type protease inhibitor) - Streptomyces sp.	53	0.98
QQBE21	Probable membrane antigen gp350 - Epstein-Barr virus	52	1.00
QQBE22	Probable membrane antigen gp220 - Epstein-Barr virus	52	1.00

Since a single "twilight" alignment can participate in many 3-alignments, the output from the program as described can be voluminous. It is possible to summarize the 3-alignments found, however, by reporting each twilight sequence that participates in a high-scoring 3-alignment only once. In this way, many 3-alignments can be output in the form of a single multiple-sequence alignment, aiding comprehension and analysis. Our general strategy for finding statistically significant 3-alignments clearly can be extended to 4-alignments, etc. However, Table 1 shows that beyond 4-alignments the power one can hope to gain is not much enhanced. Furthermore, as the number of segments in an alignment increases, the problem of whether an alignment's significance is due to only a subset of its segments becomes ever more thorny.

Biological Examples

We focus here on cases where a database search for three-way alignments reveals biologically significant sequence relationships that are not detected by the analogous pairwise database search. In two of the examples, alignments with a low degree of sequence similarity extending over 20–30 residues are found, while in one instance a short but well conserved motif is located that corresponds to an active site.

Example 1. We searched the PIR amino acid sequence database (release 22.0) with human α_1B-glycoprotein (16) (PIR code OMHU1B), a plasma glycoprotein of unknown function and member of the immunoglobulin superfamily. The immunoglobulin superfamily includes the heavy and light chains of antibodies, the histocompatibility antigens, and a wide range of receptor proteins (17). While the 2 highest-scoring database sequences from a pairwise search (Table 2) have immunoglobulin domains, the remainder of the 10 highest-scoring sequences are not members of this superfamily. Furthermore, the *P* values alone of the highest-scoring alignments are not sufficiently low to engender confidence in their biological significance. In contrast, 10 of the 11 highest-scoring similarities determined through three-way comparison are from members of the immunoglobulin superfamily, as shown in Table 3. (This table presents many distinct three-way alignments in a condensed form. Each segment shown belongs to the sequence whose PIR code appears in the second column of the table. One of the other segments in each three-way alignment belongs to the query sequence; the final

Table 3. High-scoring three-way alignments from a PIR database search with human α_1B-glycoprotein (OMHU1B)

3rd Sequence	PIR code	Start	Sequence	End	Score	*P* value
[KVMS7S 1]	B25521	21	EIVLTQSPATLSLSPGERATLSCGASQS	48	146	0.02
[B25521 21]	KVMS7S	1	DIVMTQTAPSALVTPGESVSISCRSSKS	28	146	0.02
[KVRBAH 1]	K3HUGO	2	IVLTQSPGTLSLSPGERATLSCRAAL	27	145	0.03
[K3HUGO 2]	KVRBAH	1	IVMTQTPSSKSVPVGDTVTINCQAAQ	26	145	0.03
[LVHU2 15]	S00474	31	PGEPSPPSIHPAQSELIVEAGDTLSLTCIDP	61	144	0.04
[S00474 31]	LVHU2	15	PGGSNSQTVVTQEPSLTVSPGGTVTLTCASS	45	144	0.04
[K3HUGO 1]	KVMS7A	1	DIVMTQSPTFLAVTASKKVTISCTAS	26	143	0.05
[K3HUGO 2]	KVRB29	1	IVMTQTPSSKSVPVGDTVTINCQASQ	26	143	0.05
[KVMS7S 1]	K1HUMV	1	DVQMTQSPSSLSASVGDRVIITCRASQSSVD	31	140	0.10
[K3HUVH 22]	XFSMF	10	LTMGHGNSAATVNPERAVTLNCAPTASGT	38	140	0.10
[XFSMF 10]	K3HUVH	22	IVMTQSPPTLSLSPGERVTLSCRASQSVS	50	140	0.10
Query:	OMHU1B	184	AAPPPPVLMHHGESSQVLHPGNKVTLTCVAPLSGVD	219		

PIR code	Protein description
B25521	Ig kappa chain V region 305 precursor - Human
KVMS7S	Ig kappa chain V region - Mouse
K3HUGO	Ig kappa chain V-III region - Human Gol
KVRBAH	Ig kappa chain V region - Rabbit
S00474	Kinase-related transforming protein (kit) precursor - Mouse
LVHU2	Ig lambda chain V region - Human
KVMS7A	Ig kappa chain V region - Mouse
KVRB29	Ig kappa chain V region - Rabbit
K1HUMV	Ig kappa chain V-I region - Human Mev
XFSMF	Plasminostreptin (PSTI-type protease inhibitor) - Streptomyces sp.
K3HUVH	Ig kappa chain precursor V-III region - Human Vh

Table 4. A specific three-way alignment found in a PIR database search with human α_1B-glycoprotein (OMHU1B)

PIR code	Start	Sequence	End
OMHU1B	184	AAPPPPVLMHHGESSQVLHPGNKVTLTCVAP	214
S00474	31	PGEPSPPSIHPAQSELIVEAGDTLSLTCIDP	61
LVHU2	15	PGGSNSQTVVTQEPSLTVSPGGTVTLTCASS	45

Three-way alignment score, 144; *P* value, 0.04

Sequences in pairwise alignment	Optimal score	*P* value	Score in 3-alignment
OMHU1B, S00474	58	0.52	58
OMHU1B, LVHU2	49	1.00	31
S00474, LVHU2	55	(0.88)	55

segment belongs to the sequence whose PIR code appears in the first column, along with that segment's starting position.) The eight similarities having *P* values less than 0.05 are all from sequences in the immunoglobulin superfamily. Plasminostreptin (PIR code XFSMF), which does not appear to be related to human α_1B-glycoprotein, participates in a 3-alignment with the not statistically significant *P* value 0.10.

Table 4 illustrates how three-way comparison increases the power of a database search. (Although the third pairwise alignment does not involve the query sequence, the *P* value shown is calculated using the same value of *N* used for the other two pairwise alignments.) The c-Kit sequence S00474 yielded the highest scoring pairwise alignment, with a *P* value 0.52; the other pairwise similarities that constitute this three-way alignment are even weaker. However, the optimal regions of pairwise similarity are in very good agreement, i.e., all three sequences can be aligned simultaneously with little reduction in the resulting pairwise scores. The probability of this happening is much lower than that for any of the pairwise alignments occurring by chance and consequently the *P* value of the three-way alignment is 0.04.

Example 2. The database was searched with the EbgR protein of *Escherichia coli* (18) (PIR code RPECEG); Table 5 is part of the output generated. The *ebgR* gene codes for the repressor of the EBG system of *E. coli* (evolved β-galactosidase); this operon is thought to be homologous to the *lac* operon (19). The *E. coli cyt* repressor (20, 21) (PIR code RPECCT) would have been found by a pairwise search, due to extensive similarity to the query. The pairwise similarity of the query to each of the two other sequences shown, however, is well within the realm of chance. Three-way comparison focuses attention on these sequences; in each case, the aligned region corresponds with the helix–turn–helix type of DNA-binding domain (20, 21). While the degree of sequence conservation is occasionally strong enough in the helix–turn–helix class of DNA-binding domains (22, 23) to allow detection by pairwise database searches, this example illustrates that in some cases more powerful methods are needed to detect functional domains in distantly related proteins.

Example 3. Table 6 shows the results of a database search with the protease sequence from simian AIDS retrovirus SRV-1 (24) (PIR code PRLJSA). Although a number of the relationships indicated in this alignment would be detected through a pairwise search method, this table illustrates the power of the multiple alignment approach in focusing attention on very short but highly conserved subsequences that may be critical for function, i.e., active sites. The conserved "DTG" (aspartic acid-threonine-glycine) pattern in this set of retroviral Pol proteins is the active site in aspartyl proteases. A detailed analysis of viral polymerases led to the observation of this pattern (25), yet it is apparent in a single three-way database search.

Discussion

Existing sequence database search methods have been remarkably effective in detecting biologically significant relationships. High-scoring similarities to unrelated sequences, however, can cause a researcher who uses only pairwise comparison methods to miss some biologically significant relationships. Consider the following two scenarios for evaluating the significance of a pairwise similarity.

Suppose a researcher believes a query sequence to be related to the cytochrome *c* family, and therefore performs a pairwise comparison of the query to a particular cytochrome sequence. To assess the significance of the similarity, he compares the query to 200 randomly chosen non-cytochrome sequences from the database and discovers that the query/cytochrome pair scores higher than all but one of the non-cytochrome comparisons. It is reasonable to conclude that the probability is less than 1% that the query/cytochrome relationship is random.

Suppose instead the researcher compares the query to a database of 10,000 sequences. Because of the size of the search, he might now find as many as 50 non-cytochrome sequences with scores greater than the query/cytochrome pair. Absent the prior suspicion that the query is related to

Table 5. High-scoring three-way alignments from a PIR database search with *E. coli* EbgR protein (RPECEG)

3rd Sequence	PIR code	Start	Sequence	End	Score	*P* value
[BVECPB 167]	RPECCT	11	TMKDVALKAKVSTATVSRAL	30	139	0.15
[RPECCT 11]	BVECPB	167	SQKDIAAKEGLSQAKVTRAL	186	139	0.15
[RPECCT 10]	ZBBPU2	19	TTFKQIALESGLSTGTISSFIND	41	138	0.19
Query:	RPECEG	2	ATLKDIAIEAGVSLATVSRVLND	24		

PIR code	Protein description
RPECCT	cyt repressor - Escherichia coli
BVECPB	parB protein - Escherichia coli plasmid P1
ZBBPU2	Gene B protein - Bacteriophage mu

Table 6. High-scoring three-way alignments from a PIR database search with the protease sequence from simian AIDS retrovirus SRV-1 (PRLJSA)

3rd Sequence	PIR code	Start	Sequence	End	Score	P value
[GNLJG4 132]	GNFF42	61	LLDTGADISILKENS	75	142	0.07
[GNFF42 61]	GNLJG4	132	LLDTGADDTIIKEND	146	142	0.07
[GNVWA2 63]	GNLJEV	87	KRPTTIVLINDTPLNVLLDTGADTSVL	113	135	0.33
[GNLJEV 87]	GNVWA2	63	QRPLVTIRIGGQLKEALLDTGADDTVL	89	135	0.33
[GNVWA2 63]	GNLJEW	87	KRPTTIVLINDTPLNVLLDTGADTSVL	113	135	0.33
[GNLJEV 87]	GNVWLV	63	QRPLVTIKIGGQLKEALLDTGADDTVL	89	129	0.84
[GNLJEV 87]	GNVWH3	75	QRPLVTIKIGGQLKEALLDTGADDTVL	101	129	0.84
[GNLJEV 87]	GNVWVL	72	QRPLVTIKIGGQLKEALLDTGADDTVL	98	129	0.84
[GNLJG4 131]	PNLJH2	51	ALLDTGADLTVIP	63	129	0.84
[GNFF42 61]	PNLJH1	62	LLDTGADMTVLP	73	121	1.00
[GNFF42 61]	PNLJCN	130	LLDTGADMTVLP	141	121	1.00
Query:	PRLJSA	70	QKPSLTLWLDDKMFTGLIDTGADVTIIKLED	200		

PIR code	Protein description
GNFF42	Retrovirus-related pol polyprotein (transposon 412) - Fruit fly
GNLJG4	pol polyprotein - Simian immunodeficiency virus (SIV)
GNLJEV	pol polyprotein - Equine infectious anemia virus
GNVWA2	pol polyprotein - AIDS virus ARV-2 (AIDS-associated retrovirus)
GNLJEW	pol polyprotein - Equine infectious anemia virus (clone 1369)
GNVWLV	pol polyprotein - AIDS virus LAV-1a (lymphadenopathy-associated virus)
GNVWH3	pol polyprotein - AIDS virus HTLV-III (T-cell leukemia virus, BH10)
GNVWVL	pol polyprotein - AIDS virus LV (lymphadenopathy virus)
PNLJH2	Probable protease - T-cell leukemia virus (HTLV-II)
PNLJH1	Protease - T-cell leukemia virus (HTLV-I)
PNLJCN	Protease - T-cell leukemia virus I (HTLV-I, Caribbean isolate)

cytochromes, his attention might not be drawn to this particular match on the basis of pairwise sequence similarity. A scientist with substantial experience in database searching might notice, however, several distantly related cytochromes appearing among the high-scoring sequences and investigate whether the region of similarity between the query and these cytochromes shows substantial overlap.

It is this latter approach to finding homologies that we have put into a more powerful statistical and algorithmic framework, permitting the detection of biologically significant relationships among several sequences when none of the constituent pairwise similarities are themselves statistically significant. We believe this will allow scientists who only occasionally perform protein sequence analyses to detect homologies they otherwise would have missed, and will speed and enhance studies done by more experienced researchers.

An implementation in the C programming language of the strategy described here is available from the authors upon request. The program BLAST3, which utilizes the heuristic BLAST search strategy, requires less than 4 min for a typical search of the current PIR database. It runs under either the 4.2 BSD or the AT&T System V UNIX operating system.

We appreciate valuable programming assistance from Dr. Warren Gish and helpful comments on the manuscript from Dr. David Landsman.

1. Lipman, D. J. & Pearson, W. R. (1985) *Science* **227**, 1435–1441.
2. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.*, in press.
3. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.
4. Karlin, S., Dembo, A. & Kawabata, T. (1990) *Ann. Stat.* **18**, 568–577.
5. Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 4355–4358.
6. Taylor, W. R. (1986) *J. Mol. Biol.* **188**, 233–258.
7. Patthy, L. (1987) *J. Mol. Biol.* **198**, 567–577.
8. Smith, R. F. & Smith, T. F. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 118–122.
9. Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48**, 443–453.
10. Sellers, P. H. (1974) *SIAM J. Appl. Math.* **26**, 787–793.
11. Smith, T. F. & Waterman, M. S. (1981) *Adv. Appl. Math.* **2**, 482–489.
12. Sellers, P. H. (1984) *Bull. Math. Biol.* **46**, 501–514.
13. Dayhoff, M. O., Schwartz, R. M. & Orcut, B. C. (1978) *Atlas of Protein Sequence and Structure* (Natl. Biomed. Res. Found., Washington), Vol. 5, Suppl. 3, pp. 345–352.
14. Altschul, S. F. & Lipman, D. J. (1989) *SIAM J. Appl. Math.* **49**, 197–209.
15. Coulson, A. F. W., Collins, J. F. & Lyall, A. (1987) *Computer J.* **30**, 420–424.
16. Ishioka, N., Takahashi, N. & Putnam, F. W. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 2363–2367.
17. Williams, A. F. & Barclay, A. N. (1988) *Annu. Rev. Immunol.* **6**, 381–405.
18. Stokes, H. W. & Hall, B. G. (1985) *Mol. Biol. Evol.* **2**, 478–483.
19. Valentin-Hansen, P., Larsen, J. E. L., Hojrup, P., Short, S. A. & Barbier, C. S. (1986) *Nucleic Acids Res.* **14**, 2215–2228.
20. Abeles, A. L., Friedman, S. A. & Austin, S. J. (1985) *J. Mol. Biol.* **185**, 261–272.
21. Miller, J. L., Anderson, S. K., Fujita, D. J., Chaconas, G., Baldwin, D. L. & Harshey, R. M. (1984) *Nucleic Acids Res.* **12**, 8627–8638.
22. Steitz, T. A., Ohlendorf, D. H., McKay, D. B., Anderson, W. F. & Matthews, B. W. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 3097–3100.
23. Kelley, R. L. & Yanofsky, C. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 483–487.
24. Power, M. D., Marx, P. A., Bryant, M. L., Gardner, M. B., Barr, P. J. & Luciw, P. A. (1986) *Science* **231**, 1567–1572.
25. Toh, H., Kikuno, R., Hayashida, H., Miyata, T., Kugimiya, W., Inouye, S., Yuki, S. & Saigo, K. (1985) *EMBO J.* **4**, 1267–1272.