Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 1382–1386, March 1983
Evolution

Optimal sequence alignments

(chicken hemoglobin/homology–analogy/distance similarity/gap weighting/Sellers, Needleman–Wunsch algorithms)

WALTER M. FITCH* AND TEMPLE F. SMITH†

*Department of Physiological Chemistry, University of Wisconsin, Madison, Wisconsin 53706; and †Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545

Communicated by E. Margoliash, October 15, 1982

ABSTRACT **Current theory is adequate to the task of finding an optimal alignment between two character strings such as nucleic acids. Most algorithms currently in use must fail to find the homologous alignment between a set of codons for the chicken α- and β-hemoglobin sequence when it is in fact discoverable by a more general treatment of gaps. Fundamental reasons for this are discussed.**

The importance of methods for comparing nucleotide sequences is illustrated by the fact that the first issue of *Nucleic Acids Research* in 1982 was devoted solely to computer algorithms. Most of the sequence alignment algorithms currently in use are dynamic programming algorithms along the lines originally developed in 1970 by Needleman and Wunsch (1). The rigorous mathematical methods introduced in 1974 by Sellers (2) and Wagner and Fischer (3) differ principally in that they produce an alignment minimizing a distance measure, whereas Needleman and Wunsch maximized a similarity measure. All three techniques allow for the inclusion of gaps (insertions or deletions), but their weighting (or penalty) functions considered only the existence of single sequence element gaps *per se*. In 1976, Waterman *et al.* (4) generalized the treatment of gap weights to include gaps of more than one sequence element in length, restricted only by the requirement that the weight function be monotonically increasing with the gap length. Sellers (5) then introduced the idea of separating the weight given to gaps at the ends of sequences from those in the interior of a sequence; in particular, he suggested that for many problems terminal gaps should be unweighted. Smith *et al.* (6) proved an equivalence relationship in 1981 between the Seller's metric and the Needleman–Wunsch similarity value with a gap-weighting function of the form $g + nr$, where g is the weight for the presence of the gap *per se*, r is the weight for each residue position in the gap, and n is the number of sequence elements in the gap. As noted in the *Discussion*, Dayhoff's (7) matrix bias can be shown to produce a linear gap weighting. This linear form of the gap-weighting function permits (P. Haeberli, personal communication) such an algorithm to find a solution in time proportional to the square of the sequence length rather than the cube, as would be required for weights as a nonlinear function of gap length.

In general, then, under the linear gap-weight assumption, there are at least three parameters involved in the inclusion of gaps in an optimal alignment: g and r as well as whether or not to weight one or both end (terminal) gaps. In this study, we investigated this parameter space in an attempt to understand the potential biological significance of alignments so obtained. Because the α–β hemoglobins are one of the best studied pairs of sequences involving sufficient data to allow the presumptive gaps to be known with some certainty, we have used from chicken the α and β hemoglobins for our study to see how well different procedures find the appropriate alignment.

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.

METHODS

The Sequences. The sequences chosen to be aligned were codons 40–52 (39 nucleotides) of the mRNA of chicken α hemoglobin (8) and codons 39–57 (57 nucleotides) of the mRNA of chicken β hemoglobin (9). They were chosen specifically because their alignment is believed from amino acid sequence comparisons to require two gaps as follows:

β	F-A-S-F-G-N-L-S-S-P-T-A-I-L-G-N-P-M-V
α	F-P-H-F D-L-S-H G-S-A-Q-I.

The length was deliberately kept short, partly to save computer time, but more importantly, to make the problem difficult without being intractable.

The Alignment Algorithm. The particular program used (6)‡ permits one to vary all three parameters and choose between similarity (Needleman–Wunsch) or distance (Sellers) mode. The following analyses were performed in the Needleman–Wunsch mode because, in the general case, distance algorithms cannot be defined for unweighted end gaps, an option which was to be included.

For any given set of parameter values, there may be more than one equally optimal (having identical similarity values) alignment. For example, an unmatched nucleotide at the beginning of a gap could be equally unmatched if moved to the other end of the gap. We define as a *solution set* all equally optimal alignments produced by a given set of parameter values.

The gap-weight parameter space is further degenerate in that there are *parameter domains* in the positive g/r quadrant in which all values of g and r generate the same solution set. Fortunately, the number of such domains is finite, given finite sequences.

Although the parameter space is infinite, there exists a value of g sufficiently large that no increase in the number of matching sequence elements can overcome the cost of introducing even one (weighted) gap, other than the one required when comparing sequences of unequal length with end gaps weighted.

At the other extreme, trivially small but non-zero values will obtain the solution set that maximizes the number of matches irrespective of the number of gaps required. The values of g and r should not both be set equal to zero, as those solution sets are pathological in introducing more gaps than are required to obtain the maximum number of matches. The entire parameter space is next searched in a sequential manner so that all solution-set domains are identified as outlined in the *Appendix*.

‡ In the program by Smith *et al.* in 1981, the relationship between the Sellers distance and the Needleman–Wunsch similarity was reported incorrectly. Their sum should equal α_{max}, the maximum match value, times the average length of the two sequences.

RESULTS

For unweighted terminal gaps there were 11 different solution sets. One optimal alignment from each solution set is shown in Fig. 1. A similar collection of optimal alignments for the seven solution sets for terminal gaps weighted the same as interior gaps is shown in Fig. 2. Each solution set is assigned a capital letter, which is used to label their parameter domains in Fig. 3 and 4, respectively. Solutions set A and B of the unweighted end-gap case are identical to solution sets L and M of the weighted end-gap case. An example of a pathological solution is provided by solution set Z in Fig. 2. It has no more matches at 30 than L (=A), although it has eight additional gaps involving 14 additional residues spanned by gaps. Although our program permits it, we did not explore the cases in which only one of the two end gaps would have been weighted as needed.

Solution sets A and K in Fig. 1 are the two extreme cases used in the *Appendix* to exemplify the calculation of the first provisional line separating parameter domains. Solution sets G and H were the ones obtained in testing the validity of that line. Fig. 3 shows the final result of that series. Because solution sets A and K have no common border, it is not surprising that no trace of that first provisional line remains. Forty-two of the points are required to show that the parameter domains extend to those vertices.

Alignment Q is the recognized homologous alignment except that the G alignment to the left of the first gap is assumed to be to the right of that gap if the nucleotide solution were mapped onto the amino acid solution. In the general case, it is quite reasonable that the removed triplet(s) begin and end within codons, thus simultaneously causing a neighboring nucleotide left behind to change the encoded amino acid.

The number of nucleotides that match, the number of weighted gaps, and the number of residues in weighted gaps for each optimal solution are given in Table 1. The range of the number of weighted gaps is from 0 to 10 and involves up to 22 residues spanned by weighted gaps. Note that different solution sets may have the same number of matches (G and H), the same number

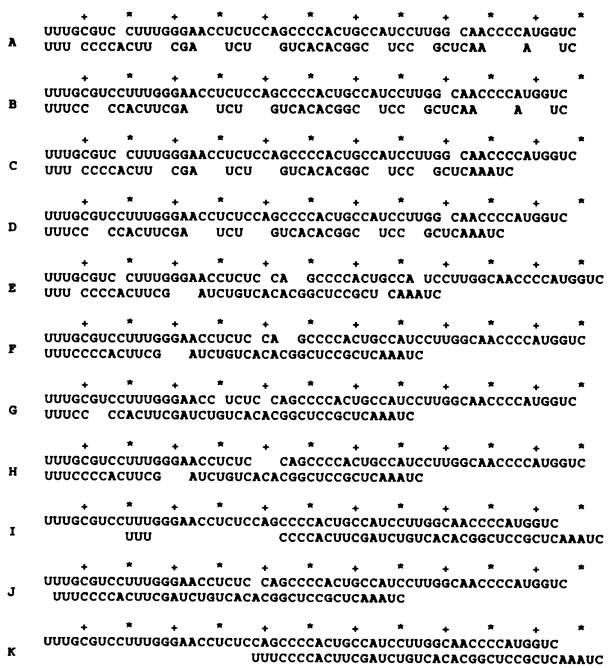

FIG. 1. Optimal nucleotide alignments with unweighted end gaps. The upper sequence is chicken β-hemoglobin mRNA, nucleotides 115–171; the lower sequence is that of chicken α-hemoglobin mRNA, nucleotides 118–156.

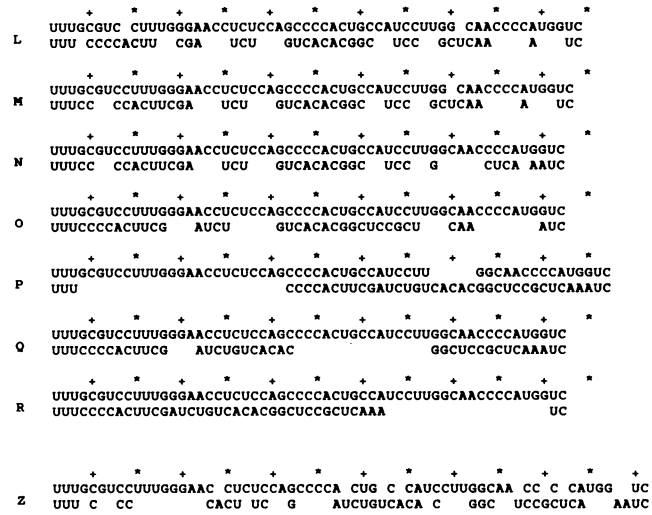

FIG. 2. Optimal nucleotide alignments with weighted end gaps. See Fig. 1 for details.

of weighted gaps (F and G), and the same number of residues in weighted gaps (F and H). A probability associated with a solution set is given also for selected cases. There is generally more than one optimal alignment. The probability attached to the solution sets are in no instance significant because of the short

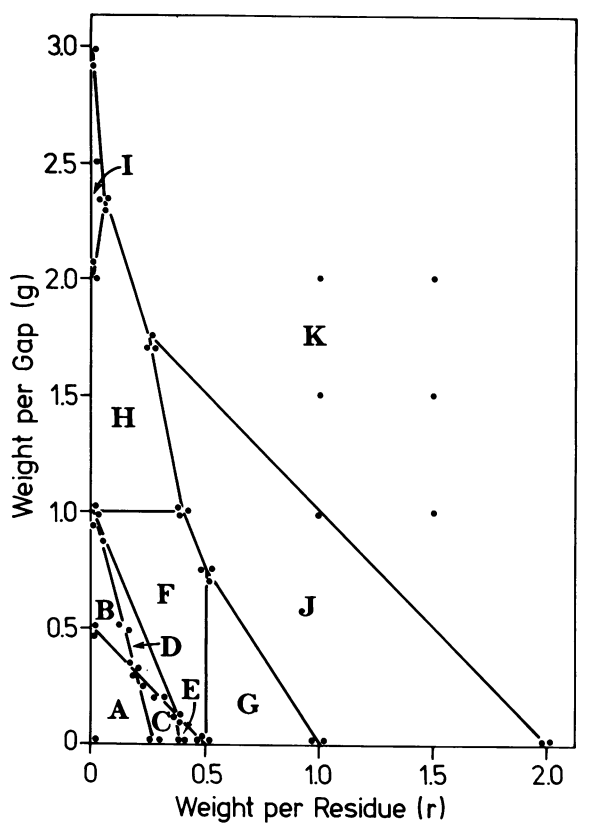

FIG. 3. Regions that give the same optimal solutions (unweighted end gaps). Weights are in nucleotides that must be subtracted from the number of nucleotide matches to give the value of an alignment. Letters correspond to the alignments in Fig. 1, and all pairs of parameters within a region give the same optimal solution set. The dots within the figure show points for which the program was run. Their locations near the vertices represent tests to assure that the limits of an optimal solution set are as stated; because region K extends to infinity with all gaps having been squeezed out, they constitute a proof that all solution sets have been found.

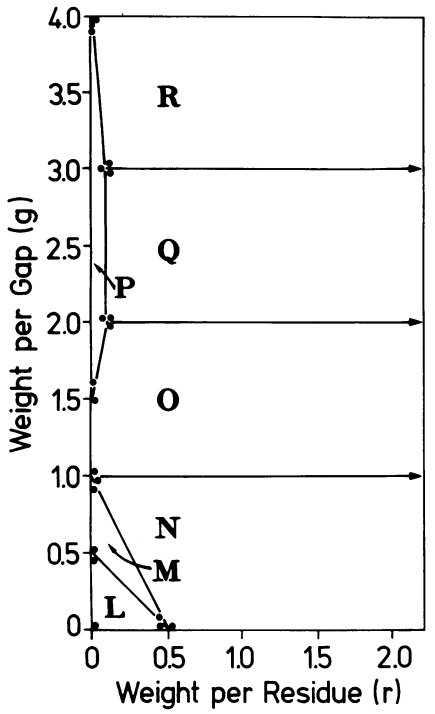

FIG. 4. Regions that give the same optimal solutions (weighted end gaps), determined as in Fig. 3 except that the alignments of Fig. 2 are used. Regions N, O, Q, and R extend rightward to infinity, and R also extends upward to infinity. This is because all four of these regions have alignments that have been reduced to the minimum of 18 residues opposite gaps, the number of nucleotides by which the longer sequence exceeds the shorter. Together with the dots near the vertices, this constitutes proof that all solution sets have been found.

sequence lengths used, but they tend to reflect the recognized underlying homology.

Table 2 shows the three possible solution sets if the same sequence is examined in terms of the encoded amino acids. In this case the match values were the maximum number of nucleotides that might possibly agree when only the pair of amino acids was known. This is three minus the common minimal base difference. The middle alignment is the known homologous alignment except that the histidine to the right of the second gap should be on the left of that gap. This arises because the histidine codon could match the leucine codon in the first and third nucleotide positions but could match the serine codon only in the third. Had the leucine been the more commonly occurring methionine of that position of other β-hemoglobin sequences, the histidine would have been properly placed. This last point shows the importance of using multiple representatives of a given sequence family.

Table 1. Characteristics of the optimal unweighted end-gap solutions

Alignment	Matches, no.	Gaps, no.	Residues in weighted gaps, no.	P
A	30	10	22	
B	29	8	20	
C	28	8	15	
D	27	6	13	
E	26	7	10	
F	24	3	6	0.47
G	23	3	4	0.64
H	23	2	6	0.55
I	21	1	14	
J	20	1	1	
K	18	0	0	0.25–0.37
L	30	10	22	0.22
M	29	8	20	
N	28	7	18	
O	25	4	18	
P	22	2	28	0.07–0.13
Q	21	2	18	0.08–0.10
R	18	1	18	0.13–0.17
Z	30	18	36	

It is of interest that the most significant alignment is the one involving the amino acid sequences ($P = 0.01$), although all such reported values here are subject to considerable uncertainty arising from there having been only 20 shuffles of the α-hemoglobin sequence in each case.

DISCUSSION

Finding Homology. There are 16 optimal solution sets to the posed alignment problem within the considered parameter space. Notice that the recognized homologous alignment Q is found only in the case of weighted end gaps and that its parameter domain, although extending rightward to infinity, does not extend to either axis. Consequently, methods that explore only along the g axis of the unweighted end-gap parameter space (Fig. 3) as does the original Needleman–Wunsch (1), or only along the r axis of the weighted end-gap parameter space (Fig. 4) as does the original Sellers (2, 10), could not find the accepted hemoglobin homologous alignment. Therefore, no such restricted method in general can be expected to identify the most likely homologous alignments. This must be the case because even the two-parameter linear gap weight is clearly not the most general one, even though it is sufficient in this rather stringent test case.

It should be noted that the bias value added by Dayhoff (7) to the similarity values of all sequence elements not aligned with gaps is equivalent to twice the r value. Because Dayhoff's gap

Table 2. Optimal amino acid alignments

	g	r	P
F-A-S-F-G-N-L-S-S-P-T-A-I-L-G-N-P-M-V F-P-H-F-D L-S-H-G-S-A-Q-I	3.01	1.50	0.07
F-A-S-F-G-N-L-S-S-P-T-A-I-L-G-N-P-M-V F-P-H-F D-L-S H-G-S-A-Q-I	2.01	1.50	0.01
F-A-S F-G-N-L-S S-P-T-A I-L-G-N-P-M-V F P-H-F D-L-S-H-G-S A-Q-I	0.01	0.50	0.09

Segments of β (upper) and α (lower) chicken hemoglobin aligned by using minimal base differences. g, Weight per gap; r, weight per residue in gap; P, probability of occurring by chance estimated from 20 shufflings of the α-hemoglobin sequence.

penalty is equivalent to the g value, internal gaps are treated by Dayhoff exactly as in this work. However, the bias method forces end gaps to be weighted by the r term even when the gap penalty is not applied to end gaps. This latter case is the one described in Ref. 7.

The superiority of this method over others in use should not obscure the fact that the gap-weighting function is not in its most general form. It is quite possible in some particularly complex case, involving perhaps several nearby deletions, that one would find only the alignment most likely to be truly homologous if the penalty for the gaps were a nonlinear, monotonically increasing function of n rather than linear. However, this will slow down the algorithm, and the best biologically appropriate form of that function is unknown at present.

Curiously, solution set Q is in the only unweighted terminal-gap domain that does not possess a portion of an edge of the parameter space. Neither D nor F do in the unweighted end-gap case, although they do share the point $g = 1.00$, $r = 0.00$ with solution sets B and H. At those precise parameter values, a good program should be able to give all alignments belonging to all four solutions sets. If only one or a few alignments are printed out, they could be peculiar to the program and its search procedure. For example, solution set Q has 13 equally optimal alignments. Notice also, that the domains, rather than being separated by the points on the boundary lines, in fact share the points on the boundary lines.

The decision as to which solution set is statistically optimal is simply a matter of taking the one with the smallest P value. Because this is only an example case, not all solution sets had a P value calculated for them, but solution set P is instructive in that it has a similar P value to that of Q, which is the recognized answer. Perhaps more scrambles would resolve the matter, but neither P is significant. If one is forced to choose, however, Q is one of only three solution sets whose gaps are all multiples of three nucleotides and, for coding sequences, gaps whose lengths are not multiples of three nucleotides induce frameshifts. These may be anticipated to be deleterious, possibly fatal, if not compensated nearby (and at essentially the same time) by other insertion/deletion events. Thus, only solution sets like Q would permit, for coding sequences, the occurrence of more than one gap to represent independent genetic events occurring at disparate times and even in different lineages.

More troubling is the matter of obtaining an honest value of P. As one increases g or r, or both, one expects the mean similarity values for the randomized sequences to decrease. Thus, for any given parameter domain, one expects the P value to be smaller for values of g and r in the upper right portion and larger for values in the lower left of the domain. This is reflected in the range of P values for solutions sets K, P, and Q, although the fact of using only 20 shuffles also contributes to the variability. The conservative approach is probably to accept P as significant only if it is determined by using g and r values in the lowest portion of their range residing in the parameter domain of the solution set being accepted.

Finding the solution set that contains the homologous alignment does not necessarily mean that that alignment will be identified by any algorithm unless (or even if) they are all printed out. In the case of alignment Q, there are 13 other equally acceptable alignments.

Nucleotide vs. Amino Acid Sequences in Testing for Similarity. One cannot completely know the nucleotide sequence of the gene from a knowledge of the amino acid sequence of its product because of the degeneracy of the genetic code. As a consequence of the smaller amount of information represented by an amino acid sequence, it would seem transparently obvious that the nucleotide sequences of a protein's gene would be superior for determining the truly homologous alignment of residues. Yet, at least statistically, the amino acid sequence alignment appears more significant in our particular case. There are a couple of reasons for this. First of all, the code is degenerate in that there are numerous nucleic acid sequences potentially encoding the same protein; secondly, most biochemical constraints of selective pressures are assumed to act at the amino acid level. This combination is related to the general principle in taxonomy and systematics that characters (nucleotide positions in this case) are unreliable for sorting out evolutionary relationships if they change their state (are substituted by another nucleotide) too rapidly because the details of the history are then overwritten. Perhaps the rate of change in the third position of codons is so great as to make those positions unreliable. The examination of the amino acid sequences may, in effect, be a filtering out of those unreliable, noisy characters, thus increasing the power of the test.

CONCLUSIONS

Methodological. (*i*) It is essential that an alignment algorithm weight both the gap and its length.

(*ii*) One should not use zero weight for either g or r, as the alignment methods will then give alignments that are adjudged equally optimal despite the fact that some of the solutions may contain more gaps or more residues spanned by gaps than do other alignments.

(*iii*) The sum of g plus r should exceed the (maximum) match value if insertions and deletions are considered to be rarer than nucleotide substitutions.

(*iv*) If a high degree of matching occurs in two portions of the sequences, the locking of the alignment in these two regions sets the minimal number of intervening residues to be spanned by gaps as the difference in residue length between those locked regions of the two sequences. Under such circumstances, the value most influential in determining the optimal alignment is g, which affects the number of gaps between those regions, rather than r which affects the number of residues gapped.

(*v*) If the end of the two sequences are not known to be matched, terminal gaps should not be weighted because the *a priori* expectation is that the ends will be unmatched. This arises not because the sequences should be presumed to be nonhomologous, but because the beginning and end points were presumably determined by procedures that had little, if anything, to do with making the ends homologous. This expectation is increased if the sequences are of unequal length to begin with. On the other hand, it would be foolish not to weight ends known to be homologous. The known homologous alignment (Q) in our test case would not have been found if terminal gaps had not been weighted. A good algorithm permits this choice.

(*vi*) The Needleman–Wunsch mode of maximizing matches is superior to the Sellers mode simply because it is feasible to allow terminal gaps to go unweighted.

(*vii*) Match (or mismatch) values other than 0 and 1 should be permitted. For example, it is reasonable to suggest that a transition difference is not as great as a transversion difference and that amino acid differences are clearly not all alike.

(*viii*) There must be a function within the program that gives one the probability that the optimal alignment might have been that good by chance.

The program was developed for the study by Smith *et al.* (6), and this study meets all of the above eight considerations and is freely available.

Interpretive. (*i*) One must distinguish between a gap and its length. Designating a gap of five residues as five gaps confuses a distinction absolutely vital to obtaining the results required.

(*ii*) One must distinguish between optimal and statistically significant. Solution set G is optimal for its parameter values, but its match value [23 − (3 gaps × 0.01 per gap) − (4 residues in gaps × 0.98 per residue in gap) = 19.05] is below the mean of 20 shuffles of the α-hemoglobin sequence as indicated by its *P* value of >0.5.

(*iii*) One must distinguish between similarity and homology. Two sequences may be similar without having a common ancestor. Ones does not "introduce gaps to optimize homology" but only to optimize similarity. Common ancestry cannot be optimized because, like pregnancy, it does not admit of degrees. Moreover, two homologous sequences in correct alignment are completely homologous even if only 50% of the aligned nucleotides match. One observes similarity and infers common ancestry. To use homology for both is to promote confusion. The meaning of homology as similarity resulting from common ancestry has 100 years of priority in biology over similarity and, in conjunction with analogy (similarity resulting from convergence), preserves an important distinction (see *v* below).

(*iv*) One must distinguish between the significance of an alignment and the correctness of the alignment. Alignment P has a *P* value as low as that of Q and could be pushed lower by adding on more codons from neighboring positions. Thus, it might be significant, but it would not be the homologous alignment except at the ends. More than half of the nucleotides are aligned nonhomologously, while those at the end that are homologous produce the low *P* value. Thus, the homology of the sequences may be reflected in a low *P* value, while the alignment is, nevertheless, largely nonhomologous.

(*v*) One must distinguish between homology and analogy. Two promoters have been demonstrated to be analogous (6). Although the degree of convergence is small, the similarity is significant and cannot be the result of homology because they are different subsequences of known homologies. Because both homology and analogy are known to occur in macromolecular sequences, it is especially hazardous to use homology for both the observation and one of two possible inferences from that observation.

APPENDIX

A sequential search of the entire parameter space is carried out by assuming that the two extreme solutions are the only two that exist, then finding the line that would divide the quadrant into their two respective domains, and, finally, showing whether the assumption was correct or not. If not, further solution sets will be uncovered that tell one how to further subdivide the quadrant.

Suppose (as in fact proves to be the case for our chicken hemoglobin sequences) that, for unweighted end gaps, the solution set with the most matches has 12 more matches, *i*, than the one with the least matches (30 vs. 18), the "improvement" being bought at the price of 10 additional internal gaps, *a*, involving 22 additional residues, *b*, spanned by gaps. These values are inserted into the formula $ag + br = i$ to give the result $10\,g + 22\,r = 12$, which will define a provisional line that intersects the *g* axis of our parameter space at 12/10 and the *r* axis at 12/22, thus dividing the quadrant into two domains that are the final answer if and only if there are no parameter values that give other solution sets.

One tests the validity of such a provisional line by choosing parameter values near its intercept with other boundaries. One might choose at one end of the line the values $g = 1.18$, $r = 0.01$ and, at the other end, $g = 0.01$, $r = 0.52$. These values result in two new solution sets with 23 matches each, but with the former having three internal gaps involving four residues and the latter having two internal gaps involving six residues. These will provide five new provisional lines that may be similarly tested. The process is repeated until all values of *g* and *r* in the immediate region where two such provisional lines intersect give the expected solution set. The details will vary with the sequences used, but the principle of dividing up the quadrant ever more finely remains constant, with the exception that the gaps need not be internal if one is weighting end gaps. A separate graph is required for weighted and unweighted end gaps.

Probability of Result. Estimating the probability that a given similarity value would arise by chance was by the common "standard measure" method. In this Monte Carlo procedure, the α-hemoglobin nucleotide sequence was randomized 20 times, and the mean and standard deviation of the 20 similarity values were calculated. The difference between the observed similarity value and the mean of the 20 scrambled cases was divided by the standard deviation to determine how many σ the observed result was above (or below) the mean to get the standard measure. The probability was taken as a one-tailed test from the normal distribution. In practice, one might shuffle the codons rather than the nucleotides if coding sequences were being examined.

We thank Ron Niece for assistance in this work and Margaret Dayhoff for her criticisms of it. This work was supported by National Science Foundation Grant DEB 78-1419T to W.M.F. and by the Los Alamos National Laboratory through its support of T.F.S.

1. Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48**, 443–453.
2. Sellers, P. H. (1974) *SIAM J. Appl. Math.* **26**, 787–793.
3. Wagner, R. A. & Fischer, M. J. (1974) *J.A.C.M.* **21**, 168–183.
4. Waterman, M. S., Smith, T. F. & Beyer, W. A. (1976) *Adv. math.* **20**, 367–387.
5. Sellers, P. H. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 3041.
6. Smith, T. F., Waterman, M. S. & Fitch, W. M. (1981) *J. Mol. Evol.* **18**, 38–46.
7. Dayhoff, M. D. (1978) *Atlas of Protein Sequence and Structure* (Nat. Biomed. Res. Found., Washington, DC), Vol. 5, Suppl. 3, pp. 6–7.
8. Deacon, N. J., Shine, J. & Haora, H. (1980) *Nucleic Acids Res.* **8**, 1187–1199.
9. Richards, R. I., Shine, J., Ullrich, A., Wells, J. R. E. & Goodman, H. M. (1979) *Nucleic Acids Res.* **7**, 1137–1146.
10. Fickett, J., Goad, W. & Kanehisa, M. (1981) *Los Alamos Sequence Library—A Data Base and Analysis System for Nucleic Acid Sequences* (Los Alamos National Laboratory, Los Alamos, NM), Rep. LA-9274.