PROTEINS: Structure, Function, and Genetics 32:88–96 (1998)

Generalized Affine Gap Costs for Protein Sequence Alignment

Stephen F. Altschul*
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

ABSTRACT **Based on the observation that a single mutational event can delete or insert multiple residues, affine gap costs for sequence alignment charge a penalty for the existence of a gap, and a further length-dependent penalty. From structural or multiple alignments of distantly related proteins, it has been observed that conserved residues frequently fall into ungapped blocks separated by relatively nonconserved regions. To take advantage of this structure, a simple generalization of affine gap costs is proposed that allows nonconserved regions to be effectively ignored. The distribution of scores from local alignments using these generalized gap costs is shown empirically to follow an extreme value distribution. Examples are presented for which generalized affine gap costs yield superior alignments from the standpoints both of statistical significance and of alignment accuracy. Guidelines for selecting generalized affine gap costs are discussed, as is their possible application to multiple alignment. Proteins 32:88–96, 1998.**

Key words: structural alignment; multiple alignment; pattern recognition; statistical significance; BRCA1

INTRODUCTION

The comparison of two protein or DNA sequences generally is guided by a similarity function that assigns a score to all possible alignments. The score for a given alignment is most often taken to be the sum of "substitution scores" for aligning pairs of residues, and "gap scores" for aligning strings of residues in one sequence with null characters introduced into the other. The gap scores in earliest common use charged a fixed penalty for each residue in either sequence aligned with a null in the other. Because under this system the cost of a gap is proportional to its length, we call these length-proportional gap costs. Using these costs, algorithms for constructing optimal global or local alignments require $O(mn)$ time, where m and n are the lengths of the sequences being compared.[1–5]

Over the years it was observed that the optimal, or highest-scoring, alignments produced by length-proportional gap costs often invoked a large number of short insertions or deletions and were not biologically plausible. That a single mutational event might insert or delete a large number of residues suggested that a long gap should not cost substantially more than a short one. The simplest way to capture this idea is to charge a gap of length k the cost $a + bk$: the existence of a gap costs a, and each residue aligned with a null costs b. In certain cases where the biologically correct alignment is known, the use of such "affine" in place of length-proportional gap costs has been shown to be necessary if the true alignment is to be the highest-scoring one.[6] Fortunately, algorithms for the construction of optimal alignments using affine gap costs are only slightly more complicated than those required for length-proportional gap costs, and require only a constant factor more space and time.[7–9] It is possible, of course, to define more complicated gap costs, for example as an arbitrary function of gap length.[10] For the class of "concave" gap costs, optimal alignment algorithms may still be constructed that require only $O(mn)$ time.[11] However, these algorithms are substantially more difficult to implement and almost all alignment programs in popular use have confined themselves to affine gap costs.

Many methods for multiple alignment, structural alignment, and sequence–structure threading have been developed. When comparing two protein structures, it is often apparent that secondary-structural elements may be superimposed closely in space, while the loops that connect them remain difficult if not impossible to align. Accordingly, many measures of the quality of a structural or threading alignment essentially ignore these intervening loops.[12–15] Similarly, many approaches to local multiple alignment confine themselves to seeking ungapped blocks of aligned residues separated by regions of variable length that are left unaligned.[16–23] One widely used database of protein motifs is constructed of just such ungapped blocks.[24] A possible view is that such constraint is imposed only for algorithmic reasons—

*Correspondence to: Stephen F. Altschul, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894. E-mail: altschul@ncbi.nlm.nih.gov
Received 5 February 1997; Accepted 22 January 1998

that fully aligning the regions between two blocks would simply be too time-consuming, and accordingly is omitted. Often this perception may be correct, but as the structural alignment example shows, it is frequently more accurate to claim that the segments separating two conserved regions should not be aligned than to impose an alignment upon them. If this is true for structural and multiple alignments, does it have any relevance for simple pairwise alignment?

One original motivation for pairwise sequence alignment was the reconstruction of molecular evolution.[3,25,26] Confining attention to substitution, insertion, and deletion mutations, one can claim that two homologous sequences have a historically correct alignment, which it is the goal of sequence comparison to approximate as well as possible. However, for distantly related proteins, an alternative viewpoint may emerge. Some protein regions are under greater structural constraint than others, and therefore evolve more slowly. As a result, two proteins may share several regions with recognizable similarity, separated by regions bearing no detectable mutual relationship. At the structural level, this lack of similarity may actually reflect a loss of three-dimensional correspondence. Even from an evolutionary perspective, while it may still make sense to align these regions, the requisite information for doing so simply may have been lost.

This article introduces a generalization of affine gap costs, applicable to both global and local pairwise alignments, that within a larger alignment permits apparently unrelated sequence regions to remain unaligned. For local alignments, the distribution of optimal alignment scores is shown empirically to follow an extreme value distribution. The relevant statistical parameters may be estimated for different gap cost settings. The effectiveness of a given alignment scoring system may be measured both by the degree to which it yields statistically significant scores for related sequences, and by the degree to which the optimal alignments it generates conform to biological reality. In many cases, generalized affine gap costs prove superior to traditional costs by both of these criteria. Empirical studies can guide gap cost selection,[27] but general considerations regarding features of the alignments sought also can inform the choice. Generalized affine gap costs may be introduced into applications that employ pairwise sequence alignment, such as progressive multiple alignment.

GAPS AND THEIR ASSOCIATED COSTS

Traditionally, a gap within a pairwise alignment is defined to consist of k residues from a single sequence, and affine gap costs assign it a score of negative $a + bk$. We generalize the notion of gap to involve k_1 residues from sequence A and k_2 residues from sequence B. One can assign a cost to such a gap in many different ways. Perhaps the simplest extension of affine gap costs gives this gap a score of negative $a + b(k_1 + k_2)$.[12] However, this definition is indifferent between, say, 30 residues gapped out of a single sequence and 15 residues left unaligned in each of the two sequences. From structural and perhaps even evolutionary considerations, one may wish to prefer the latter case. Accordingly, we introduce a three-parameter generalization of affine gap costs, in which the score $-a$ is assessed for the existence of a gap, $-b$ for each residue inserted or deleted, and $-c$ for each pair of residues left unaligned. More formally, the score for a gap involving k_1 and k_2 residues, with $k_1 \geq k_2$, is negative $a + b(k_1 - k_2) + ck_2$. We will represent these generalized affine gap costs by the ordered triple (a, b, c). When $c = \infty$, these costs reduce to traditional affine gap costs, and when $c = 2b$, they reduce to those proposed by Zuker and Somorjai.[12] Note that we have adopted a different parameter-naming convention than on occasion is used elsewhere.[27,28] Specifically, the gap opening score $-a$ is sometimes taken to include the score for the first inserted or deleted residue, while here it is not.

Generalized affine gap costs may be used in either the global or local alignment context. For global alignments, one has as always the choice of whether to score end gaps differently than internal gaps. Standard dynamic programming algorithms for either global or local alignment can easily accommodate generalized affine gap costs, in an analogous manner to their treatment of traditional affine gap costs. The main difference is that once a gap has been opened, diagonal moves within the path graph are permitted, with a score $-c$, in addition to vertical or horizontal moves with a score $-b$ (Fig. 1). Implementations that return the optimal alignment score and a representation of all optimal alignments require $O(mn)$ time and space.[7,8] If only a single optimal alignment is required, the space requirement can be reduced to $O(min(m, n))$.[9] The details of these algorithms are easily reconstructed, and will be omitted here.

LOCAL ALIGNMENT STATISTICS

Little is known concerning the distribution of optimal global alignment scores from the pairwise comparison of random sequences. In contrast, the random distribution of optimal local alignment scores is quite well understood. The prototypical case is that of local alignments in which gaps are forbidden; the scores of such alignments have been shown analytically to follow an extreme value distribution.[29,30] Given a matrix of substitution scores s_{ij} for aligning pairs of residues, and background probabilities p_i for the occurrence of residues within the sequences, the values of two key parameters, λ and K, may be calculated. (The expected score $\Sigma_{i,j}\, p_i p_j s_{ij}$ for aligning two random residues must be negative

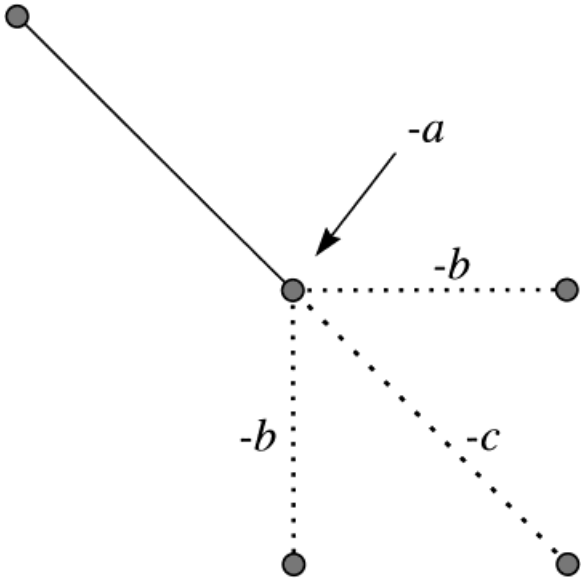

Fig. 1. A schematic representation of how scores are assessed within a path graph when generalized affine gap costs are employed. The score $-a$ is charged for the existence of a gap; $-b$ for each unpaired residue left unaligned; and $-c$ for each pair of residues left unaligned. The solid diagonal line represents a pair of aligned residues; the dotted diagonal line, a pair of unaligned residues; and the dotted horizontal and vertical lines, single unaligned residues.

for the theory to hold). Using these parameters, the raw score S of the optimal local alignment may be converted to a normalized score S' by the formula

$$S' = \frac{\lambda S - \ln K}{\ln 2}. \qquad (1)$$

Such a normalized score S' is said to be expressed in *bits*. The expected number of distinct segment pairs with normalized score greater than or equal to x is then well approximated by the formula

$$E(S' \geq x) \approx N/2^x \qquad (2)$$

where the search-space size N is the product of the lengths of the sequences being compared.[29,30]

Once gaps and their associated costs are allowed within local alignments, the statistical theory outlined above is no longer known to hold. However, some theory[31] and many computational experiments[28,32,33] strongly suggest that it does. The only practical difference is that one may no longer calculate λ and K analytically. Instead, they must be estimated by either random simulation or the comparison of real but unrelated sequences.[28,32–35]

All statistical studies of gapped local alignments to date, of course, have employed at most affine gap costs. While it appears likely that the same statistical theory will apply to the scores of alignments generated using the generalized affine gap costs introduced here, it is nevertheless desirable to adduce some empirical support. Accordingly, we generated 24,000 pairs of length 1,000 random protein sequences, using the background amino acid frequencies of Robinson and Robinson.[36] Each pair was compared using a scaled version (Fig. 2) of the BLOSUM-62 amino acid substitution matrix,[37] and (120, 10, 3) generalized affine gap costs. A histogram of the 24,000 optimal local alignment scores produced is shown in Figure 3. The best fit of an extreme value distribution[38] to these data was estimated by the maximum likelihood method,[39] and the resulting curve is shown in Figure 3. A χ^2 goodness-of-fit test, with 275 degrees of freedom, had the value 290.1; a worse fit would be expected 27% of the time even were the extreme value theory precisely valid. Analogously to traditional affine gap costs,[28] we have performed more extensive tests on generalized affine gap costs (data not shown) to establish that they conform to other aspects of the basic statistical theory.[29,30]

To employ Equations (1) and (2), all that is needed are estimates of λ and K. For any set of gap costs we consider, these parameters were estimated as described above. The standard error for the resulting estimate of λ was approximately 0.5%, and for K approximately 5%. However, the method for estimating these parameters[39] has the effect of making their errors approximately proportional. As a result, the standard error for normalized scores in the range of 40 bits is about 0.1 bits. In addition to being subject to stochastic error, the parameter estimates are of course dependent on the particular random protein model used. With λ and K in hand, Equation (1) converts raw scores into normalized scores, expressed in bits. This normalization permits the alignment scores generated by different substitution matrices and gap costs to be directly compared.[40,41]

BIOLOGICAL EXAMPLES

For sequences that are closed related, generalized affine gap costs will provide no advantage to traditional gap costs, because related regions will not be interrupted by regions without detectable similarity. To study whether generalized gap costs can improve the detection of weak relationships, we used an appropriately modified version of the Smith-Waterman algorithm[4] to search release 34 plus updates of the SWISS-PROT database[42] with 11 protein queries. For homologous database sequences that barely attained statistical significance, we compared the scores returned by generalized and traditional gap costs. The results are shown in Table I.

For our database searches, we used a version of the BLOSUM-62 amino acid substitution matrix[37] (Fig. 2), scaled by a factor of 10 so that gap scores could be kept integral. To select reasonable gap costs for general-purpose sequence comparison, there is little substitute for empiricism. An exhaustive evalu-

```
A   39
R  -14  55
N  -15  -4  57
D  -18 -16  13  58
C   -4 -34 -27 -35  86
Q   -8  10   0  -3 -29  53
E   -9  -1  -3  15 -36  19  49
G    2 -23  -4 -13 -25 -18 -21  56
H  -16  -2   6 -11 -30   4  -1 -20  75
I  -13 -30 -32 -31 -12 -28 -32 -37 -32  40
L  -15 -22 -34 -36 -13 -21 -28 -36 -28  15  38
K   -7  21  -2  -7 -30  13   8 -15  -7 -27 -24  45
M   -9 -14 -22 -31 -14  -4 -20 -27 -16  11  20 -14  54
F  -22 -28 -30 -35 -24 -32 -32 -31 -12  -2   4 -31   0  60
P   -8 -21 -20 -15 -28 -13 -11 -21 -22 -28 -29 -10 -25 -36  74
S   11  -8   6  -3  -9  -1  -1  -3  -9 -23 -24  -2 -15 -24  -8  39
T    0 -11   0 -11  -9  -7  -9 -16 -17  -7 -12  -7  -7 -21 -11  14  45
W  -25 -27 -37 -42 -23 -19 -28 -25 -23 -26 -16 -30 -14   9 -37 -28 -24 105
Y  -18 -17 -21 -31 -24 -14 -20 -30  17 -13 -11 -18 -10  29 -29 -17 -16  22  66
V   -2 -25 -29 -31  -8 -22 -24 -31 -31  25   8 -23   7  -8 -23 -16  -1 -28 -12  38

     A   R   N   D   C   Q   E   G   H   I   L   K   M   F   P   S   T   W   Y   V
```

Fig. 2. A scaled version of the BLOSUM-62 amino acid substitution matrix.[37] Because we wish to consider gap costs that would be fractional in the usual units in which that matrix is expressed, and so that we may continue to deal in integers, we have multiplied the standard matrix by 10. Since the matrix was originally constructed by rounding real numbers to the nearest integer, we have returned to the raw data to gain precision.

ation of gap cost parameter space for the most sensitive parameter settings is beyond the scope of this article. However, we have found that in conjunction with our scaled BLOSUM-62 matrix, (120, 10, 3) gap costs prove generally effective; the corresponding statistical parameters are estimated at $\lambda \approx 0.0286$ and $K \approx 0.041$. For the purposes of comparison, it is appropriate to select a set of traditional affine gap costs ($c = \infty$) with nearly identical λ. This renders raw alignment scores nearly comparable, allowing the comparison to hinge almost completely on the differential scoring of gaps. Accordingly, we kept the b gap cost parameter fixed at 10, and lowered the penalty a for the existence of a gap from 120 to 97; the statistical parameters for (97, 10, ∞) gap costs were then estimated to be $\lambda \approx 0.0286$ and $K \approx 0.046$. Are these a reasonable set of traditional gap costs to employ? Pearson[27] has conducted performance tests on a large variety of search algorithms, substitution matrices, and traditional affine gap costs. In his nomenclature our scoring system would correspond roughly to (11, 1) gap costs used in conjunction with the standard BLOSUM-62 matrix. While Pearson offers no single prescription of scoring system for database searching, this one at least falls within the set of reasonable choices.

To focus on distantly diverged but homologous sequences, we analyzed only alignments that appeared moderately significant ($0.1 > E > 0.0001$) using at least one of our two sets of gap costs. For our 11 queries, the number of alignments satisfying this condition ranged from 1 to 71. SWISS-PROT annotation suggested that all such alignment represented biologically meaningful relationships, with the exception of one returned by the histocompatibility antigen query. This single false positive received a score greater by 0.9 bits using the traditional gap costs. As shown in Table I, for eight of the 11 queries the mean normalized score using (120, 10, 3) gap costs was higher than that with (97, 10, ∞) gap costs. Averaged over queries, the mean score differential was 0.6 bits, corresponding to a factor of 1.5 in statistical significance. The optimal score for most database sequences is not greatly affected by the use of one set of gap costs or the other. Nevertheless, for eight (vs. two) queries, the use of generalized gap costs improved at least one alignment score by more than three bits, enough to affect materially the ability to recognize a similarity as statistically significant.

One may inquire into not only the score of a sequence similarity, but also the accuracy of the alignment to which it corresponds, as measured by some gold standard, such as the alignment's congruence with a multiple or a structural alignment. It is not clear how one is best to construct such an objective standard. We took the relatively straightforward approach of applying one iteration of the PSI-BLAST program[43] to each of our queries. This program constructs a multiple alignment from the significant alignments returned by an initial data-

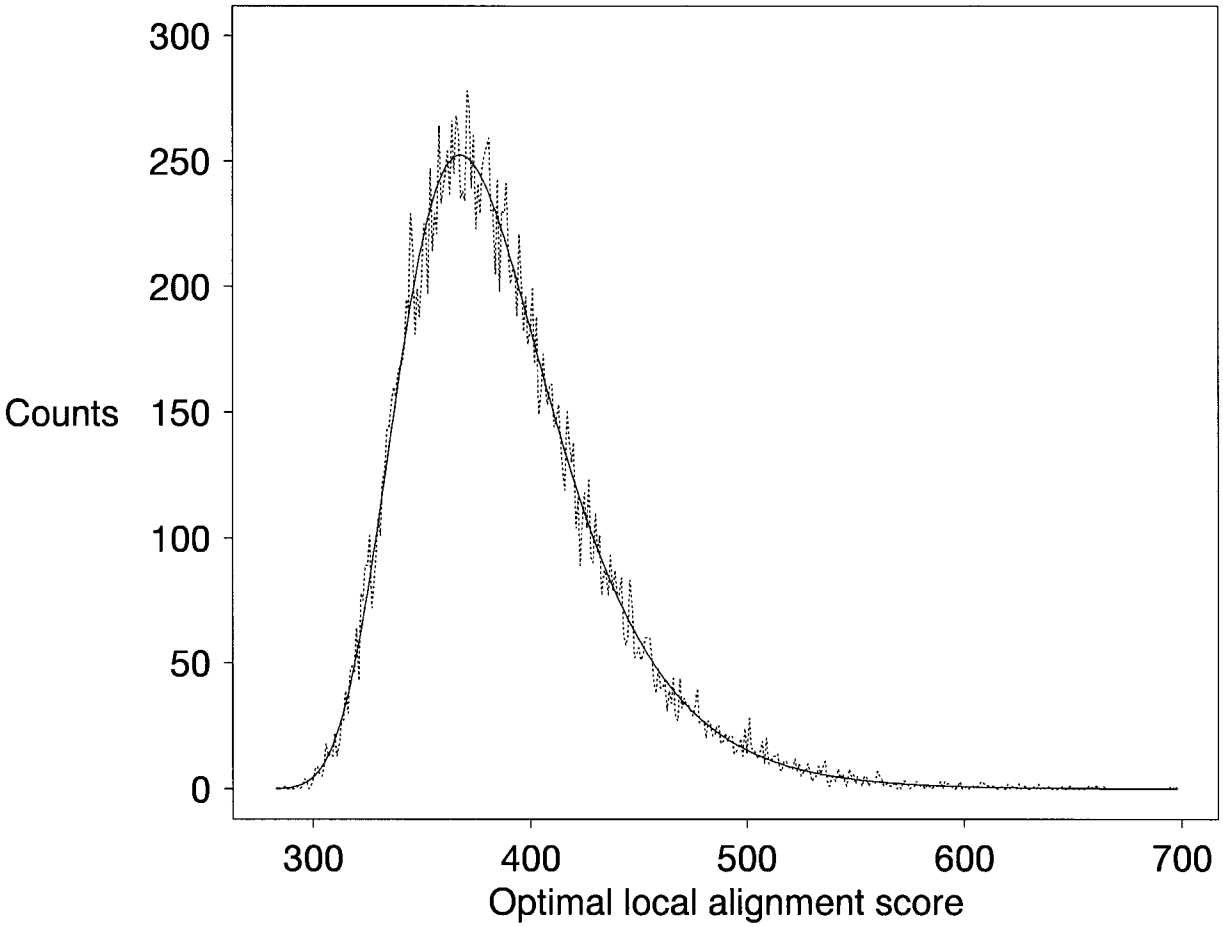

Fig. 3. A histogram of the optimal local alignment scores of 24,000 pairs of random sequences of length 1,000, generated using the amino acid frequencies of Robinson and Robinson.[36] Scores were calculated using the BLOSUM-62 substitution matrix of Figure 2, and (120, 10, 3) gap costs. The superimposed extreme value distribution[38] was calculated to fit the data by the method of maximum likelihood.[39] A χ^2 goodness-of-fit test, with 275 degrees of freedom, has the value 290.1.

base search, and then uses a position-specific score matrix derived from this alignment to perform a subsequent search. So that reasonable credence could be given to the alignments used to construct PSI-BLAST's score matrix, we employed a stringent initial cutoff *E* value of 10^{-10}. Also, we ran PSI-BLAST using traditional gap costs, which should tend to bias the alignments it returns in favor of the pairwise alignments produced by the same costs. Of the 22 similarities of Table I for which generalized gap costs produced substantially greater scores, 19 (120, 10, 3)-alignments conformed better than did (97, 10, ∞)-alignments to the corresponding PSI-BLAST alignments, two equivalently, and one worse (alignment results not shown). Conversely, of the four similarities for which traditional gap costs produced substantially greater scores, one (97, 10, ∞)-alignment conformed better to the corresponding PSI-BLAST alignment, two equivalently, and one worse. This asymmetrical result suggests that generalized affine gap costs, in addition to returning higher scores for moderately similar sequences, also tend to produce alignments that conform better to biological reality.

To illustrate the potential utility of generalized affine gap costs, we consider the conserved domain that is shared by the BRCA1 protein,[44] the human p53-binding protein 53BP1,[45] and many other human, yeast, and even bacterial proteins involved in cell cycle checkpoints.[46–48] Using the 202-residue, putatively globular, C-terminal domain of BRCA1 as the query in a database search, the original clue to the existence of this superfamily was an alignment with 53BP1. With the default scoring system provided by the BLAST program,[49] the alignment in isolation was not statistically significant, and only subsequent motif searches and multiple alignments established the relationship.[46] (For the original database search performed, the search-space size was approximately 1.2×10^{10}, implying that a normalized score of 37.8 bits was necessary for statistical significance.) Here, we compare the C-terminal domain of BRCA1 with 53BP1 using both sets of gap

TABLE I. Relative Sensitivity of Traditional and Generalized Affine Gap Costs*

Protein family	SWISS-PROT accession number of query	Number of moderately significant alignments	Alignments whose score is greater by at least three bits when using (97, 10, ∞) gap costs	Alignments whose score is greater by at least three bits when using (120, 10, 3) gap costs	Average improvement in score (bits) yielded by (120, 10, 3) gap costs
Serine protease	P00762	7	0	1	−0.1
Serine protease inhibitor	P01008	2	0	1	1.9
Ras	P01111	23	0	1	0.5
Globin	P02232	56	0	8	1.2
Hemagglutinin	P03435	4	0	0	0.5
Interferon α	P05013	1	0	0	1.7
Alcohol dehydrogenase	P07327	10	0	0	0.4
Histocompatibility antigen	P10318	71	3	2	−0.8
Cytochrome P450	P10635	20	0	4	0.6
Glutathione transferase	P14942	15	1	1	−0.4
H^+-transporting ATP synthase	P25705	13	0	4	1.3

*Using a generalization of the Smith-Waterman algorithm,[4] all queries were compared to release 34 plus updates of the SWISS-PROT database[42] (68,619 sequences; 24,728,649 amino acids). Alignment scores were derived from the scaled BLOSUM-62 matrix of Figure 2 and both traditional (97, 10, ∞) and generalized (120, 10, 3) affine gap costs. *E* values were calculated for both sets of gap costs using Equations (1) and (2). An edge-effect correction[28] for search-space size was employed, based on a calculated relative entropy[41] of 0.65 bits for ungapped alignments. Moderately significant alignments are defined as those whose smaller *E* value is <0.1 and whose larger *E* value is >0.0001.

costs described above. The optimal score yielded by (97, 10, ∞) gap costs is 34.3 bits, while that yielded by (120, 10, 3) gap costs is 38.5 bits; alignments achieving these scores are shown in Figure 4A and B. The score of the latter result is greater by 4.2 bits, corresponding to a factor of 18 in statistical significance. Furthermore, the alignment of Figure 4B nearly agrees with that implied by the multiple alignment of Koonin et al.,[46] while the alignment of Figure 4A diverges substantially (and presumable inaccurately) over its central region. This poorly conserved region is left substantially unaligned by the generalized affine gap costs; notice that one pair of segments remains unaligned in Figure 4B even though alignment could be imposed without introducing null characters into either sequence. Tellingly, once other sequences are added to the alignment, these segments span a region into which gaps must be introduced.[46]

FURTHER THOUGHTS ON GAP COST SELECTION

Because for a given set of substitution costs it has not been easy to define the optimal gap costs, one approach that has been advocated is to try them all. It can be shown that the space defined by the gap cost parameters may be divided systematically into regions in which the same alignments are optimal. Parametric alignment programs that perform such a dissection of parameter space have been described and made available.[51,52] One problem with this approach is that it generates a potentially very large number of alignments, with no guidance for choosing among them. Normalized scores, however, can provide an objective criterion for choosing among parameter settings.[40] The problem with applying them to parametric alignment is that the boundaries of parameter–space regions can not be predicted beforehand, and the stochastic experiment required to estimate λ and *K* with any accuracy for a single set of parameters requires many minutes of computational time on a standard current workstation.

An alternative approach is to precompute λ and *K* for many points placed regularly through a reasonable region of gap cost space. One may then simply calculate the optimal alignment score for each gap cost setting, and return those costs and the associated alignment that yield the highest normalized score, and thus the most significant result. One disadvantage is that there is no guarantee that the preselected gap cost settings include ones that are even near optimal for the problem at hand. Furthermore, it must be recognized that, while one may use the normalized score of Equation (1) as an objective criterion for selecting a set of gap costs, it is improper to use Equation (2) to calculate an *E* value from the normalized score. The reason is that one has performed multiple tests, and optimized among them.[40] One may calculate a conservative upper bound on the *E* value by multiplying that derived from Equation (2) by the number of parameter sets examined, but, due to the high degree of correlation among tests, this generally yields a gross overestimate. However, if the same sets of gap costs are to be examined repeatedly, it is possible but laborious to estimate the parameters for the new extreme value distribution that results from optimizing over the normalized scores.[40]

```
(a)

BRCA1  1699 RTLKYFLGIAGGKWVVSYFWVTQSIKERKMLNEHDFevrgdVVNGRNHQGPKRAR 1753
            RT KYFL +A G   VS+ WV  S    ++ N  ++     ++ +      +R
53BP1   866 RTRKYFLCLASGIPCVSHVWVHDSCHANQLQNYRNY-----LLPAGYSLEEQRIL  915

                  ########################################
BRCA1  1754 ESQDRKifrgleiccyGPFTNMP----TDQLEWMVQLC-------GASVVKELSS 1797
            + Q R+           PF N+     +DQ +  ++L        GA+ VK+  S
53BP1   916 DWQPRE----------NPFQNLKvllvSDQQQNFLELWseilmtgGAASVKQHHS  960


BRCA1  1798 FT----LGTGVHPIVVVQPDAwteDNGFHAIGQMCEAPVVTREWVL 1839
                  ++ GV  +VV  P             ++  + PVV++EWV+
53BP1   961 SAhnkdIALGVFDVVVTDPSC---PASVLKCAEALQLPVVSQEWVI 1003


(b)

BRCA1  1699 RTLKYFLGIAGGKWVVSYFWVTQSIKERKMLNEHDFevrgdvvngrnhqgpKRAR 1753
            RT KYFL +A G   VS+ WV  S    ++ N  ++                +R
53BP1   866 RTRKYFLCLASGIPCVSHVWVHDSCHANQLQNYRNYllpagyslee-----QRIL  915


BRCA1  1754 ESQDRKi-FRGLEIccygpftnmptdqlewmvqlcGASVVKELSSft----LGTG 1803
            + Q R+  F+ L++                      GA+ VK+  S      ++ G
53BP1   916 DWQPREnpFQNLKVllvsdqqqnflelwseilmtgGAASVKQHHSsahnkdIALG  970


BRCA1  1804 VHPIVVVQPDawtedngfhaiGQMCEAPVVTREWVL 1839
            V  +VV  P            ++  + PVV++EWV+
53BP1   971 VFDVVVTDPScpasvlkc---AEALQLPVVSQEWVI 1003
```

Fig. 4. Two alignments of the C-terminal domain of human breast cancer type 1 susceptibility protein (BRCA1; SWISS-PROT accession number P38398), and a fragment of the human p53 binding protein 1 (53BP1; GenBank[50] accession number U09477). Upper case is used for aligned residues and lower case for gapped residues. **a:** The optimal local alignment, using (97, 10, ∞) gap costs, has score 34.3 bits (raw score 723). Pound signs indicate a region where this alignment diverges substantially with that below. **b:** The optimal local alignment, using (120, 10, 3) gap costs, has score 38.5 bits (raw score 822).

Are there any theoretical considerations that can guide the choice of gap costs? Recall that even when no insertions or deletions need to be invoked, generalized affine gap costs may leave unaligned a diverged region that separates two related ones. One may calculate the approximate minimum length such a region needs to have before leaving it unaligned becomes profitable. First, from the substitution matrix used and the background amino acid frequencies, the expected score $-s$ for aligning two random residues may be calculated. Then for each pair of unrelated residues left unaligned, one gains on average a score of $s - c$. However, to realize this gain, one must pay a gap opening penalty of a. Thus, on average, it is beneficial to leave unaligned two unrelated segments when they are of length at least $l = a/(s - c)$. (Of course if a gap needs to be introduced in any case due to an insertion or deletion, it pays to leave any contiguous, diverged segments unaligned.) For the matrix of Figure 2, s is approximately 10, so for (120, 10, 3) generalized affine gap costs, $l \approx 17$. It is evident from this analysis that one of the main reasons for using generalized affine gap costs is substantially lost if c is greater than s, or even quite close to it.

When the score $-X$ for a region within an alignment is sufficiently negative, it generally makes more sense to break the alignment into two separate ones.[28,53] If one imagines the end of a given pair of aligned segments to be fixed, it is then possible to define the maximum extent of a gap that imposes a cost less than X. The shape, within a path graph, of this allowable gap region will depend on the relative values of the gap cost parameters b and c (Fig. 5), while its size will depend additionally on a and the nominal score X. For example, using (120, 10, 3) gap costs, with $\lambda \approx 0.0286$, for a gap to impose a penalty of fewer than 15 bits it may have a nominal cost no greater than 363. If no residue pairs are left unaligned, the maximal number of inserted or deleted residues is then 24, while if no residues are inserted or deleted, the maximal number of unaligned pairs is 81. Given a sense of the maximal desirable extent of

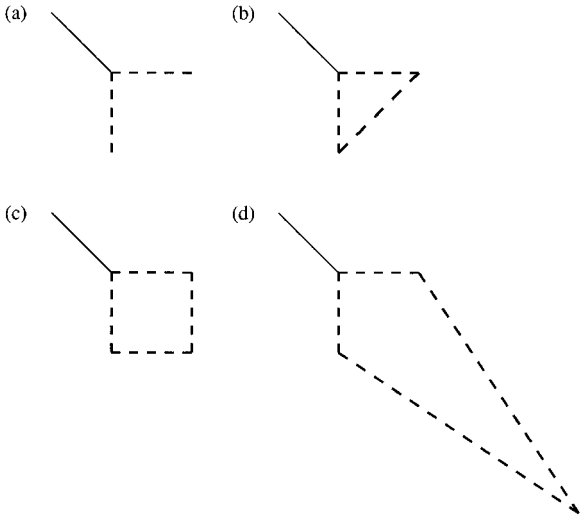

Fig. 5. The greatest extent, within a path graph, of a gap with cost less than X when (**a**) $c = \infty$, (**b**) $c = 2b$, (**c**) $c = b$, and (**d**) $c = b/3$. Given a string of aligned residues, represented by a solid diagonal line, the line representing the next string of aligned residues must start within the region enclosed by dashed lines.

a gap, one may be guided by such calculations in one's choice of gap costs.

CONCLUSION

We have seen a number of cases in which generalized affine gap costs improve somewhat the ability to detect biological relationships, as well as to construct biologically accurate alignments. Whether these costs should be incorporated into database search programs such as Fasta,[54] or gapped versions of BLAST,[28,43] depends on whether the slight increase in sensitivity is deemed worth the slight decrease in speed. For a program such as PSI-BLAST,[43] however, generalized affine gap costs may offer a more substantial improvement, because better alignment accuracy in the output from one database search can engender a more sensitive position-specific score matrix for the next.

There are other sequence comparison formalisms into which generalized affine gap costs might be incorporated. Many multiple alignment programs depend on a progressive alignment strategy, in which at first two and then greater numbers of sequences are coalesced into a single alignment.[55–63] One difficulty with this approach is that alignments formed early in the process are constructed in ignorance of most of the available data, and therefore may easily freeze in a mistake. By permitting poorly conserved regions to be left unaligned, generalized affine gap costs may partially mitigate this problem. However, extending generalized affine gap costs to multiple alignments undoubtedly will entail unforeseen technical difficulties, both definitional and algorithmic.[64] Also, the increasingly studied Hidden Markov Model formalism for representing protein families[65–68] may be able to subsume all that generalized affine gap costs can offer to the multiple alignment problem.

As the original motivation for generalized affine gap costs suggests, nonaligned regions may correspond to loops separating modular or secondary structural elements. The literature on protein secondary- and tertiary-structure prediction is too large to be reviewed here. However, ideas roughly corresponding to the one studied above are already in common use (e.g., Ref. 14). Thus it is unlikely that generalized affine gap costs have much to offer structural analysis. Because the field of pairwise sequence comparison has been fairly thoroughly plowed, one is accustomed to trying to generalize its ideas to multiple and structural alignment. It is worth recognizing that here the generalization has proceeded in the opposite direction.

ACKNOWLEDGMENTS

I thank Drs. David Lipman, Eugene Koonin, Charles Lawrence, and Stephen Bryant for helpful discussions, and Dr. Thomas Madden for programming assistance.

REFERENCES

1. Needleman, S.B., Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol. 48:443–453, 1970.
2. Sankoff, D. Matching sequences under deletion–insertion constraints. Proc. Natl. Acad. Sci. U.S.A. 69:4–6, 1972.
3. Sellers, P.H. On the theory and computation of evolutionary distances. SIAM J. Appl. Math. 26:787–793, 1974.
4. Smith, T.F., Waterman, M.S. Identification of common molecular subsequences. J. Mol. Biol. 147:195–197, 1981.
5. Sellers, P.H. Pattern recognition in genetic sequences by mismatch density. Bull. Math. Biol. 46:501–514, 1984.
6. Fitch, W.M., Smith, T.F. Optimal sequence alignments. Proc. Natl. Acad. Sci. U.S.A. 80:1382–1386, 1983.
7. Gotoh, O. An improved algorithm for matching biological sequences. J. Mol. Biol. 162:705–708, 1982.
8. Altschul, S.F., Erickson, B.W. Optimal sequence alignment using affine gap costs. Bull. Math. Biol. 48:603–616, 1986.
9. Myers, E.W., Miller, W. Optimal alignments in linear space. Comput. Appl. Biosci. 4:11–17, 1988.
10. Waterman, M.S., Smith, T.F., Beyer, W.A. Some biological sequence metrics. Adv. Math. 20:367–387, 1976.
11. Miller, W., Myers, E.W. Sequence comparison with concave weighting functions. Bull. Math. Biol. 50:97–120, 1988.
12. Zuker, M., Somorjai, R.L. The alignment of protein structures in three dimensions. Bull. Math. Biol. 51:55–78, 1989.
13. Greer, J. Comparative modeling of homologous proteins. Meth. Enzymol. 202:239–252, 1991.
14. Bryant, S.H., Lawrence, C.E. An empirical energy function for threading protein sequence through the folding motif. Proteins 16:92–112, 1993.
15. Lathrop, R.H., Smith, T.F. Global optimum protein threading with gapped alignment and empirical pair score functions. J. Mol. Biol. 255:641–665, 1996.
16. Sobel, E., Martinez, H. A multiple sequence alignment program. Nucl. Acids Res. 14:363–374, 1986.
17. Posfai, J., Bhagwat, A.S., Posfai, G., Roberts, R.J. Predictive motifs derived from cytosine methyltransferases. Nucl. Acids Res. 17:2421–2435, 1989.
18. Lawrence, C.E., Reilly, A.A. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins 7:41–51, 1990.

19. Smith, H.O., Annau, T.M., Chandrasegaran, S. Finding sequence motifs in groups of functionally related proteins. Proc. Natl. Sci. U.S.A. 87:826–830, 1990.
20. Leung, M.Y., Blaisdell, B.E., Burge, C., Karlin, S. An efficient algorithm for identifying matches with errors in multiple long molecular sequences. J. Mol. Biol. 221:1367–1378, 1991.
21. Schuler, G.D., Altschul, S.F., Lipman, D.J. A workbench for multiple alignment construction and analysis. Proteins 9:180–190, 1991.
22. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., Wootton, J.C. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. Science 262:208–214, 1993.
23. Tatusov, R.L., Altschul, S.F., Koonin, E.V. Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks. Proc. Natl. Acad. Sci. U.S.A. 91:12091–12095, 1994.
24. Henikoff, S., Henikoff, J.G. Automated assembly of protein blocks for database searching. Nucl. Acids Res. 19:6565–6572, 1991.
25. Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C. A model of evolutionary change in proteins. In: "Atlas of Protein Sequence and Structure," Vol. 5, Suppl. 3, Dayhoff, M.O. (ed.). Washington: National Biomedical Research Foundation, 1978:345–352.
26. Schwartz, R.M., Dayhoff, M.O. Matrices for detecting distant relationships. In: "Atlas of Protein Sequence and Structure," Vol. 5, Suppl. 3, Dayhoff, M.O. (ed.). Washington: National Biomedical Research Foundation, 1978:353–358.
27. Pearson, W.R. Comparison of methods for searching protein sequence databases. Prot. Sci. 4:1145–1160, 1995.
28. Altschul, S.F., Gish, W. Local alignment statistics. Meth. Enzymol. 266:460–480, 1996.
29. Karlin, S., Altschul, S.F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. U.S.A. 87:2264–2268, 1990.
30. Dembo, A., Karlin, S., Zeitouni, O. Limit distribution of maximal nonaligned two-sequence segmental score. Ann. Prob. 22:2022–2039, 1994.
31. Arratia, R., Waterman, M.S. A phase transition for the score in matching random sequences allowing deletions. Ann. Appl. Prob. 4:200–225, 1994.
32. Smith, T.F., Waterman, M.S., Burks, C. The statistical distribution of nucleic acid similarities. Nucl. Acids Res. 13:645–656, 1985.
33. Waterman, M.S., Vingron, M. Sequence comparison significance and Poisson approximation. Stat. Sci. 9:367–381, 1994.
34. Collins, J.F., Coulson, A.F.W., Lyall, A. The significance of protein sequence similarities. Comput. Appl. Biosci. 4:67–71, 1988.
35. Mott, R. Maximum likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. Bull. Math. Biol. 54:59–75, 1992.
36. Robinson, A.B., Robinson, L.R. Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins. Proc. Natl. Acad. Sci. U.S.A. 88:8880–8884, 1991.
37. Henikoff, S., Henikoff, J.G. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U.S.A. 89:10915–10919, 1992.
38. Gumbel, E.J. "Statistics of Extremes." New York: Columbia University Press, 1958.
39. Lawless, J.F. "Statistical Models for Lifetime Data." New York: John Wiley & Sons, Inc., 1982:141–202.
40. Altschul, S.F. A protein alignment scoring system sensitive at all evolutionary distances. J. Mol. Evol. 36:290–300, 1993.
41. Altschul, S.F. Amino acid substitution matrices from an information theoretic perspective. J. Mol. Biol. 219:555–565, 1991.
42. Bairoch, A., Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. Nucl. Acids Res. 26:38–42, 1998.
43. Altschul, S.F., Madden, T.L., Schäffer, A.A., et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucl. Acids Res. 25:3389–3402, 1997.
44. Miki, Y., Swensen, J., Shattuck-Eidens, D., et al. A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. Science 266:66–71, 1994.
45. Iwabuchi, K., Bartel, P.L., Li, B., Marraccino, R., Fields, S. Two cellular proteins that bind to wild-type but not mutant p53. Proc. Natl. Acad. Sci. U.S.A. 91:6098–6102, 1994.
46. Koonin, E.V., Altschul, S.F., Bork, P. BRCA1 protein products: Functional motifs. Nature Genet. 13:266–268, 1996.
47. Bork, P., Hofmann, K., Bucher, P, Neuwald, A.F., Altschul, S.F., Koonin, E.V. A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. FASEB J. 11:68–76, 1997.
48. Callebaut, I., Mornon, J.-P. From BRCA1 to RAP1: A widespread BRCT module closely associated with DNA repair. FEBS Lett. 400:25–30, 1997.
49. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. Basic local alignment search tool. J. Mol. Biol. 215:403–410, 1990.
50. Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J. Ouellette, B.F.F. GenBank. Nucl. Acids Res 26:1–7, 1998.
51. Gusfield, D., Balasubramanian, K., Naor, D. Parametric optimization of sequence alignment. In: "Proceedings of the Third Annual ACM-SIAM Symposium on Discrete Algorithms." New York: ACM Press, 1992:432–439.
52. Waterman, M.S., Eggert, M., Lander E. Parametric sequence comparisons. Proc. Natl. Acad. Sci. U.S.A. 89:6090–6093, 1992.
53. Karlin, S., Altschul, S.F. Applications and statistics for multiple high-scoring segments in molecular sequences. Proc. Natl. Acad. Sci. U.S.A. 90:5873–5877, 1993.
54. Pearson, W.R., Lipman, D.J. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U.S.A. 85:2444–2448, 1988.
55. Waterman, M.S., Perlwitz, M.D. Line geometries for sequence comparisons. Bull. Math. Biol. 46:567–577, 1984.
56. Bains, W. MULTAN: A program to align multiple DNA sequences. Nucl. Acids Res. 14:159–177, 1986.
57. Barton, G.J., Sternberg, M.J. A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons. J. Mol. Biol. 198:327–337, 1987.
58. Feng, D., Doolittle, R.F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. 25:351–360, 1987.
59. Taylor, W.R. Multiple sequence alignment by a pairwise algorithm. Comput. Appl. Biosci. 3:81–87, 1987.
60. Corpet, F. Multiple sequence alignment with hierarchical clustering. Nucl. Acids Res. 16:10881–10890, 1988.
61. Hein, J. Unified approach to alignment and phylogenies. Methods Enzymol. 183:626–645, 1990.
62. Berger, M.P., Munson, P.J. A novel randomized iterative strategy for aligning multiple protein sequences. Comput. Appl. Biosci. 7:479–484, 1991.
63. Higgins, D.G., Sharp, P.M. Clustal V: Improved software for multiple sequence alignment. Comput. Appl. Biosci. 8:189–191, 1992.
64. Altschul, S.F. Gap costs for multiple sequence alignment. J. Theor. Biol. 138:297–309, 1989.
65. Tanaka, H., Ishikawa, M., Asai, K., Konagawa, A. Hidden Markov models and iterative aligners: Study of their equivalence and possibilities. In: "Proceedings of the First International Conference on Intelligent Systems for Molecular Biology." Hunter, L., Searls, D., Shavlik, J. (eds.). Menlo Park, CA: AAAI Press, 1993:395–401.
66. Baldi, P., Chauvin, Y., Hunkapiller, T., McClure, M.A. Hidden Markov models of biological primary sequence information. Proc. Natl. Acad. Sci. U.S.A. 91:1059–1063, 1994.
67. Krogh, A., Brown, M., Mian, I.S., Sjölander, K., Haussler, D. Hidden Markov models in computational biology: Applications to protein modeling. J. Mol. Biol. 235:1501–1531, 1994.
68. Eddy, S.R., Mitchison, G., Durbin, R. Maximum discrimination hidden Markov models of sequence consensus. J. Comput. Biol. 2:9–23, 1995.