BIOINFORMATICS

Vol. 18 no. 11 2002
Pages 1500–1507

# Empirical determination of effective gap penalties for sequence comparison

*J. T. Reese and W. R. Pearson**

*Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908, USA*

Received on January 4, 2002; revised on April 9, 2002; accepted on April 18, 2002

## ABSTRACT

**Motivation:** No general theory guides the selection of gap penalties for local sequence alignment. We empirically determined the most effective gap penalties for protein sequence similarity searches with substitution matrices over a range of target evolutionary distances from 20 to 200 Point Accepted Mutations (PAMs).

**Results:** We embedded real and simulated homologs of protein sequences into a database and searched the database to determine the gap penalties that produced the best statistical significance for the distant homologs. The most effective penalty for the first residue in a gap $(q + r)$ changes as a function of evolutionary distance, while the gap extension penalty for additional residues $(r)$ does not. For these data, the optimal gap penalties for a given matrix scaled in 1/3 bit units (e.g. BLOSUM50, PAM200) are $q = 25 - 0.1 \bullet$ (target PAM distance), $r = 5$. Our results provide an empirical basis for selection of gap penalties and demonstrate how optimal gap penalties behave as a function of the target evolutionary distance of the substitution matrix. These gap penalties can improve expectation values by at least one order of magnitude when searching with short sequences, and improve the alignment of proteins containing short sequences repeated in tandem.

**Contact:** wrp@virginia.edu

## INTRODUCTION

Sequence similarity searching and sequence alignment programs have become indispensable tools for biologists. These programs are routinely used to identify homologous sequences, to infer the structure and function of proteins and even to analyze patterns in entire genomes or proteomes.

Alignment algorithms typically allow residues in one sequence to be aligned to a gap in the other sequence in exchange for a penalty. Most algorithms employ an affine gap penalty scheme, in which a penalty $q$ is accessed for the existence of a gap and another usually smaller penalty $r$ is accessed for extending the gap (Waterman *et al.*, 1976; Gotoh, 1982; Fitch and Smith, 1983). Thus the penalty for the entire gap is $q + r \bullet k$, where $k$ is the gap length.

### Gap penalties as log-odds ratios

Although the statistical behavior of local alignment scores is well understood both for ungapped (Karlin and Altschul, 1990) and gapped (Mott, 1992; Altschul *et al.*, 1997; Pearson, 1998) local alignments, current models provide little guidance in the selection of gap penalties. For very distant relationships, experience suggests that gap penalties that are as low as possible, but still produce local alignments between unrelated sequences, are the most effective (Vingron and Waterman, 1994; Pearson, 1995, 1998).

For alignments between more closely related sequences, an information theoretic perspective can provide guidance in selecting gap penalties. Altschul (1991) has shown that in the context of local sequence alignment, the values in residue substitution matrices can be considered 'log-odds ratios'. That is, an entry in the matrix is the ratio of the probability of an amino acid $i$ aligning with amino acid $j$ $(q_{ij})$ because they diverged from a common ancestor and the probability of the two amino acids $i$ and $j$ being aligned due to chance $(p_i \bullet p_j)$. From this perspective, $q_{i-}$, the probability of an amino acid $i$ being inserted into or deleted from a protein, might be estimated from multiple alignments. However, $(p_i \bullet p_-)$, the probability of aligning residue $i$ to a gap by chance, is difficult to estimate because the background frequency of gaps $(p_-)$ is unknown and probably changes with different gap penalty values.

Consequently, gap penalty parameters have been optimized empirically based on their performance in similarity searches (Pearson, 1995), accuracy of the alignment generated (Fitch and Smith, 1983; Vogt *et al.*, 1995), or maintenance of expected statistical characteristics for unrelated sequences (Vingron and Waterman, 1994). These methods typically have focused on one or a few substitution matrices, e.g. BLOSUM50 or PAM250, but the results are often extrapolated to use with all substitution matrices with the same scale (λ, Altschul, 1991). Thus, the gap penalties used by default in the FASTA

*To whom correspondence should be addressed.



© Oxford University Press 2002

program[†], $q = 10$ and $r = 2$, maximize sensitivity at long evolutionary distances (Pearson, 1995, 1998). If the target frequency for indels ($q_{i-}$) is lower for short evolutionary distances, the gap penalties should be higher. Thus, 'distant' gap penalties are probably too low for short evolutionary distances, just as scoring matrices with long target distances are inefficient at identifying closely related sequences (Altschul, 1991). The loss of efficiency from low gap penalties is mitigated by the fact closely related sequences are easy to find and relatively easy to align, so their homology probably will be evident even if inappropriate gap penalties are used. The loss of efficiency at shorter evolutionary (target) distances with inappropriate ('deep') scoring matrices is most noticeable when short sequences are aligned (Altschul, 1991); gap penalty choice is more likely to be crucial with short sequences as well. Higher gap penalties with shallow scoring matrices should reduce the number of gaps allowed between the query and unrelated sequences, resulting in lower scores for unrelated sequences and improved expectation values and bit scores for homologs. Moreover, improved gap penalties should improve alignments of short tandemly repeated subsequences.

Here we empirically determine effective gap penalties for scoring matrices targeted at different evolutionary distances. Our results from simulated sequences and real sequences agree well, and provide an empirical basis for the selection of gap penalties for a given substitution matrix. These improved gap penalties improve sensitivity when searching with short sequences and also improve the quality of the alignments.

## SYSTEM AND METHODS

We determined the gap penalties that maximized the effectiveness of sequence alignment programs in finding homologs of a given sequence. The homologs were either real or simulated using evolutionary models described previously.

### Generation of simulated sequences

Twenty sequences were selected from 20 different protein families from the SWISSPROT protein database release 34 and reversed to yield a pseudo-random sequence. Artificial homologs of a given sequence were generated using `seqevolver` (Reese *et al*., in prep.; http://fasta.bioch.virginia.edu/seqevolver). Briefly, point mutations were generated after the Dayhoff model (Dayhoff *et al.*, 1978). The PAM-x probability matrix was generated by matrix multiplication of the appropriate number of PAM1 probability matrices. A given amino acid was mutated by selecting a random number from a uniform distribution and mutating to the amino acid corresponding to the row in which the cumulative sum of probabilities was greater than the random number. Indels were generated using the model of Benner *et al.* (1993). Given a PAM distance $d$, the probability of an indel at any amino acid is:

$$P(\text{indel}) = 0.0224 - 0.0219\text{e}^{-0.01168 \bullet d}$$

and the length of the indel is independent of PAM distance, and is selected randomly from a Zipfian distribution with an exponent of 1.7 (Benner *et al.*, 1993).

### Selection of real sequences

Protein sequences with 50 homologs at a given PAM distance were chosen from the SWISSPROT protein database. PAM distances between the sequences and their homologs were measured by aligning the query to each homolog using the Smith-Waterman algorithm, BLOSUM50 substitution matrix and gap penalties of $q = 10$ and $r = 2$, and obtaining the percent identity. PAM distance was inferred from percent identity *id* using

$$d = -83 \bullet \ln(1 - ((100 - id)/100)/0.83)$$

where $d$ is distance in PAM units (Gu and Zhang, 1997). PAM distances did not change significantly when re-measured using a substitution matrix with a target distance equal to the average PAM distance of the 50 homologs (data not shown). Thirty-five sets of homologs were used, representing 12 protein families and a range of distances from 20 to 180 PAMs. The average of the standard deviations of PAM distances between sequences and their 50 homologs for the 35 sets of homologs was 3.94 PAMs. Additional information about the sequences used in this analysis can be accessed at http://www.people.virginia.edu/~wrp/papers/gappen/supplement.html.

### Database searches

Synthetic or true homologs were embedded in a database of 23 981 unrelated protein sequences. For the reversed synthetic sequences, the database was an annotated subset of SwissProt 34; the real protein sequences were embedded in 23 981 reversed SwissProt 34 sequences. The database was searched using either FASTA, ktup = 2 or SSEARCH with a substitution matrix corresponding to the average PAM distance between the query sequence and its 50 homologs. Bit scores and expectation values were calculated empirically by FASTA/SSEARCH (Pearson, 1998; Pearson and Wood, 2001).

[†] Previous versions of the FASTA programs implement gap penalties as the penalty for the first residue in a gap ($q+r$) and the penalty for each additional residue ($r$). Other programs, such as BLAST, refer to a gap open penalty ($q$) and an extension penalty ($r$). For this paper, we use the more widely accepted gap-open/gap-extend values, and the current version of FASTA (version 3.4) has been modified to use open/extend penalties and to incorporate the gap penalties described in this work.

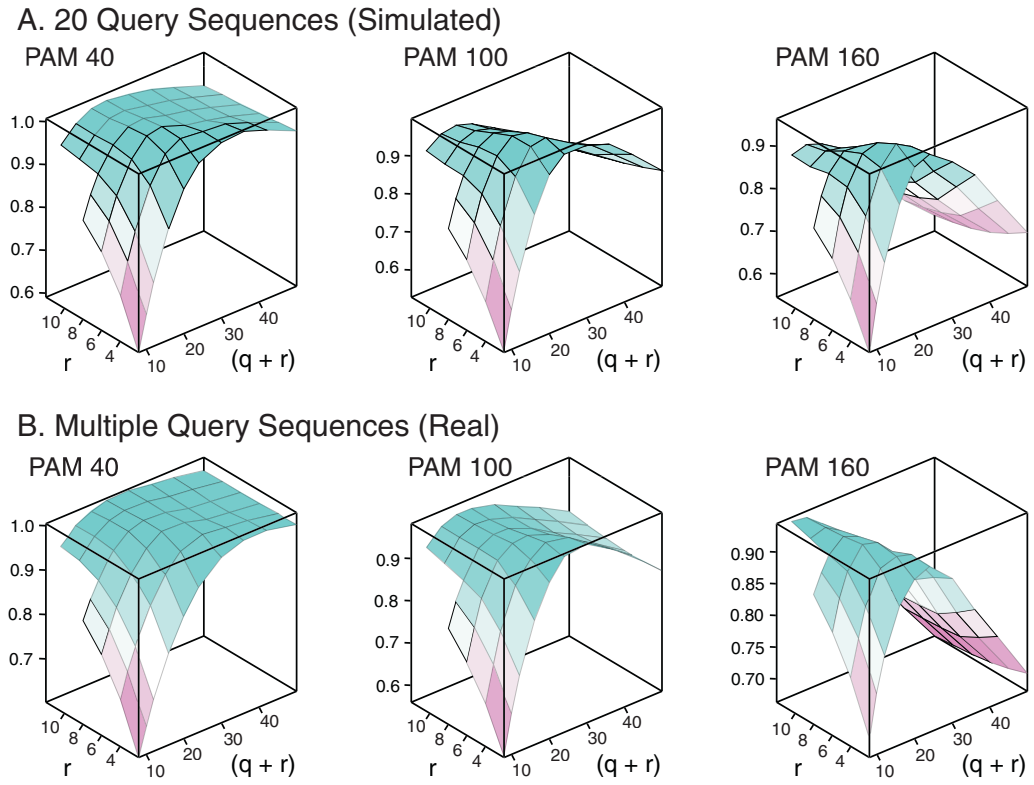


**Fig. 1.** Performance of gap penalties at different PAM distances. A database containing 50 simulated (A) or real (B) homologs separated from the query sequence by the given PAM distance was searched using SSEARCH (Pearson, 2000) with a substitution matrix whose target distance corresponded to the given PAM distance. A. Performance of $(q + r)$ and $r$ values at different PAM distances using simulated sequences. For each set of homologs, the median of the bit scores for the lowest scoring five homologs was determined. The results from the 20 sets of homologs were combined using a local regression as described in Systems and Methods. The normalized bit score for each $(q + r)$ and $r$ combination is plotted. B. Performance of $(q + r)$ and $r$ values at different PAM distances using real sequences. As in A, the results for each set of homologs and PAM distance were normalized to 1 and combined using a local regression as described in System and Methods. The normalized bit score for each $(q + r)$ and $r$ combination is plotted.

### Measurement of optimal gap penalties

We used the median bit scores of the five lowest scoring of the 50 homologs to measure gap penalty performance. Bit scores were used instead of expectation values because the expectation values were not measurably larger than zero for the most closely related sequences. When results from several sets of homologs at a given PAM distance were combined (Figure 1), the results for each homolog set were normalized to a value from 0 to 1, with 1 being the highest bit score for the homolog set. The results for that PAM distance were then combined using local regression (`loess` function of S-PLUS version 5). When $(q + r)$ or $r$ was considered as a function of PAM distance (e.g. Figure 2), the optimal value for either $(q + r)$ or $r$ for each of the 35 sets of homologs was determined. Linear regressions were performed using the `lm()` function of S-PLUS version 5. In the case of ties (multiple gap penalties all yielding the highest bit score for the median of the lowest scoring five homologs), the $(q + r)$ or $r$ values were weighted according to the inverse of the number of values participating in the tie; otherwise, each $(q + r)$ or $r$ was assigned a weight of 1.

### Substitution and probability matrices

Amino acid replacement frequencies $q_{ij}$ and background frequencies were obtained from amino acid exchange data generated from SWISSPROT release 15.0 (Jones *et al.*, 1992). The CALCPAM program (Jones *et al.*, 1992) was used to calculate log-odds scoring matrices at different evolutionary distances.

## RESULTS

The probabilities of replacement mutations $q_{ij}$ increase with evolutionary distance. For example, the probability $(q_{ij})$ of finding aspartic acid 'D' and cysteine 'C' aligned in two homologous proteins after 200 Point Accepted

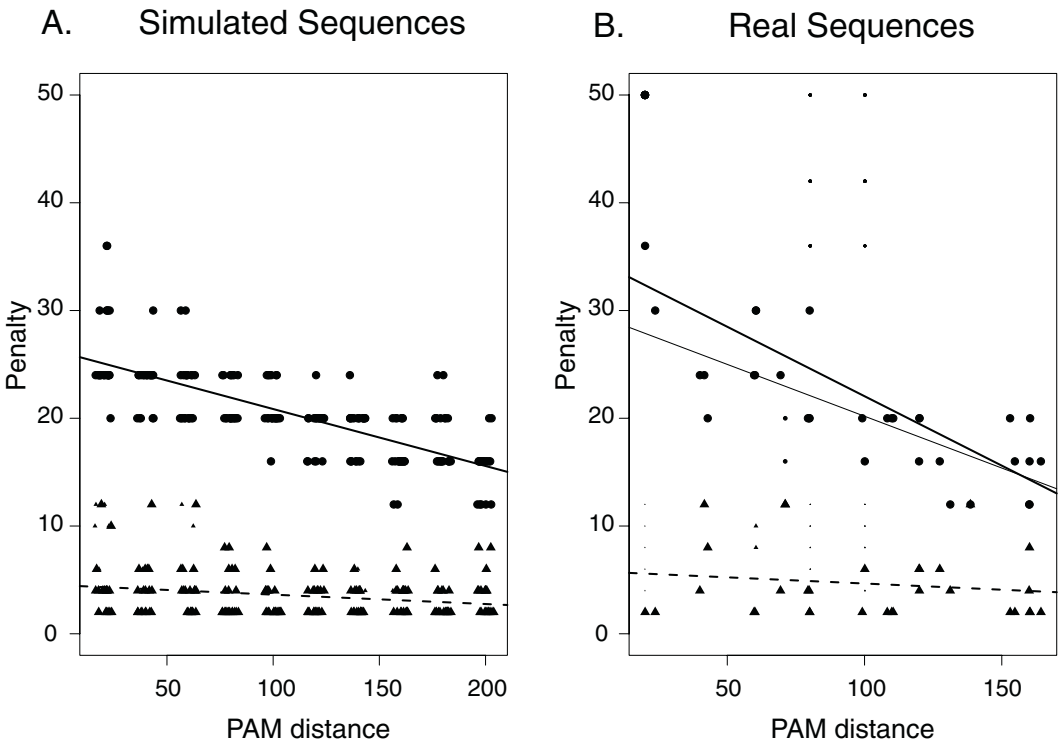


**Fig. 2.** Change in optimal gap penalties with PAM distance. A. Optimal gap penalties for simulated sequence data. The optimal $(q + r)$ (circles) or $r$ (triangles) for each of 20 sets of homologs at each of ten PAM distances (from PAM 20 to PAM200 at intervals of 20 PAMs) is shown. A small amount of scatter was introduced into the x-axis to show overlapping symbols. The least squares regressions for both $(q + r)$ (solid line) and $r$ (dashed line) are shown ($r^2 = 0.55$, $F$-statistic $= 254.5$, $p = 0$ and $r^2 = 0.05$, $F$-statistic $= 11.86$, $p \leqslant 0.0007$ for $(q + r)$ and $r$ regressions, respectively); character size indicates weight in the regression (see System and Methods). B. Optimal gap penalties for real sequence data. For each of the sets of homologs at a given PAM distance, the $(q + r)$ (circles) or $r$(triangles) at which the median of the bit scores of the lowest scoring five homologs was maximal is plotted. The least squares regression lines for $(q + r)$ (darker solid line) and $r$ (dashed line) are shown ($r^2 = 0.39$, $F$-statistic $= 47.12$, $p \leqslant 10^{-9}$ and $r^2 = 0.02$, $F$-statistic $= 1.76$, $p \leqslant 0.18$ for $(q + r)$ and $r$ regressions, respectively); character size indicates weight in the regression (see System and Methods). Lighter solid line indicates the regression for $(q + r)$ when $(q + r)$ values from outlier points from three homolog sets are omitted ($r^2 = 0.59$, $F$-statistic $= 46.36$, $p \leqslant 10^{-7}$).

Mutations (PAMs, 200 mutations per 100 residues) is 0.00052, which is over 50 times greater than the $(q_{ij})$ after only ten PAMs ($1.02 \times 10^{-5}$). We hypothesized that the target frequency of indels should increase with evolutionary distance as well, and consequently optimal gap penalties should decrease with increasing target evolutionary distance. Conversely, optimal gap penalties for shorter evolutionary distances should be higher.

We chose an empirical approach to determining optimal gap penalties, using both simulated and real sequence data. We embedded homologs in a database of unrelated sequences, and determined the gap penalties that yielded the highest bit scores for distantly related homologs. Maximizing the bit score with respect to this measure maximizes the sensitivity of detection of distant homologs while still allowing the detection of closer homologs. Results are presented in terms of optimal first residue penalty $(q + r)$ and additional residue penalty $r$ instead of $q$ and $r$ because the correlation between evolutionary distance and $(q + r)$ is much stronger than the correlation with $q$.

As expected, search performance depends strongly on gap penalties. For one homolog alignment at 100 PAMs, gap penalties of $(q + r) = 20$, $r = 4$ resulted in a bit score of approximately 165, while gap penalties of $(q + r) = 12$, $r = 2$ resulted in a bit score of approximately 80. This 85 bit increase in similarity score corresponds to a $2^{85}$ ($\sim 10^{25}$) fold increase in statistical significance. For this 216 amino acid query sequence, the average similarity increased from 0.37 bits per residue to 0.76 bits per residue. For a search of a database the size of SWISSPROT, a score of 37 bits is required to

obtain a statistical significance $E() < 10^{-3}$. In this case, the more effective gap penalties decreased the sequence length required to obtain a score of 37 bits from 100 amino acids to 49 amino acids. Thus, proteins containing a single immunoglobulin domain (68 amino acids) or a single fibronectin type-3 domain (86 amino acids) at an evolutionary distance of 100 PAMs might be missed with the lower gap penalties, but should be easily detected with the gap penalties we describe.

Search effectiveness is more sensitive to the choice of $(q + r)$ than to $r$; bit scores vary much more with $(q + r)$ than with $r$, especially near the best gap penalties. For example, at 160 PAMs, bit scores ranged from 40 to 65 when $(q + r)$ is varied and $r$is held constant at 4; in contrast, bit scores range only from about 65 to 68 when $r$ is varied and $(q + r)$ is 18. When the results for 20 different sets of simulated homologs were combined, the same trend was observed (Figure 1a). Similar results were seen when real sequences were used to evaluate the gap penalties (Figure 1b).

The optimal $(q+r)$ values decreased from 25 to 15 with increasing PAM distance (Figure 2a). The most effective extension penalty was relatively constant at $r = 5$ (while the decrease in optimal $r$ with respect to PAM distance was statistically significant, $p \leqslant 0.0007$, the slope was very slight, 0.0087; Figure 2a). We confirmed this result using real sequence data from a diverse set of protein families. Again, the optimal values for $(q + r)$ decreased with PAM distance. A very slight decrease in optimal $r$ with respect to PAM distance was observed, but this decrease was not statistically significant ($p \leqslant 0.18$) (Figure 2b); the optimal value for $r$ was approximately 5. Compared with the simulated data, the slope of the fitted line for $(q + r)$ with respect to PAM distance (Figure 2b, darker solid line) was steeper and the $y$-intercept was slightly higher than for the simulated data. Because the residuals from three sets of points (at 20, 80 and 100 PAM units, Figure 2) were inordinately high, the regression was redone excluding these outliers (Figure 2b, lighter solid line); the resulting robust regression agrees better with the simulated data for optimal $(q + r)$, although both the $y$-intercept and slope were still slightly greater than the corresponding values for the simulated data. In general, the results from the real and simulated sequence data agree well.

The results from simulated and real sequence data provide an empirical basis for the selection of effective gap penalties. Taken together, these data indicate the most effective first residue penalty $(q + r)$ changes appreciably as a function of PAM distance, while the most effective subsequent residue penalty $r$ stays nearly constant with PAM distance in the range relevant to target distances of PAM matrices. Specifically, the robust regression using real sequence data (Figure 2b) indicates that the most effective gap penalties for a matrix in 1/3 bit units (e.g. BLOSUM50, PAM200) are $q = 25 - 0.1 \bullet$ (target PAM distance), $r = 5$. For matrices scaled in 1/2 bit units (e.g. matrices at distances of PAM120 or less used by BLAST), the formula is $q = 16.7 - 0.067 \bullet$ (target PAM distance), $r = 3$.

To confirm that the above results may be extrapolated to heuristic methods based on the Smith–Waterman algorithm (Altschul *et al.*, 1997; Pearson, 2000), we repeated the above experiments using FASTA instead of the SSEARCH. As before, the most effective $(q + r)$ value decreased appreciably as a function of PAM distance, while the most effective $r$ value did not (data not shown). The values for $q$ and $r$ implied by these data were very similar ($q = 22 - 0.081 \bullet$ (target PAM distance) and $r = 6$) to those for the Smith–Waterman experiments. This suggests that the gap penalties implied by Smith-Waterman results can be extrapolated to heuristic algorithms such as FASTA.

The gap penalties suggested by these data are considerably larger than those used by default in many sequence alignment programs, especially for matrices with short target PAM distances. For example, for a matrix in 1/3 bit units with a target distance of 20 PAM units, these data suggest gap penalties of approximately $q = 23, r = 5$, compared with the default values for older versions of FASTA of $q = 10, r = 2$. Likewise, current WWW versions of BLAST (2.2.2) provide the PAM30 matrix with gap open penalties ranging from 5–10 and gap extension penalties of 1 or 2; our results suggest that a gap open penalty of 14–15 and an extension penalty of 3 would be more effective.

These gap penalties improve sensitivity when one must use shallow matrices, such as when one is searching with short sequences. We evaluated this increase in sensitivity by examining the expectation values for each of 59 homolog pair alignments, where one of the sequences contained only 20 amino acids. When searching with gap penalties of $q = 10$ and $r = 2$ (the default for older versions of FASTA), the homolog was found ($E() < 10^{-3}$) 33 of 59 times (Figure 3). When then same experiment was done using improved gap penalties of $q = 21$ and $r = 5$, the homolog was found 39 of 59 times, an 18% improvement in sensitivity. Moreover, for alignments that fell near the threshold of sensitivity ($E() \sim 10^{-2}$) the expectation value improved 10–100-fold when the most effective gap penalties were used (Figure 3). In two cases sensitivity decreased, probably in cases where large indels occurred in the 20 amino acid sequence or the corresponding homologous sequence to which it was aligned. In the large majority of cases, the expectation value improved, indicating that the gap penalties suggested here improve sensitivity when searching with short sequences.

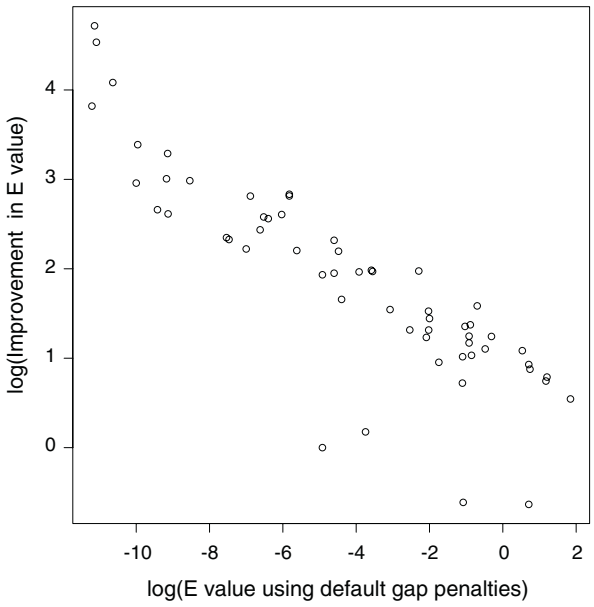


**Fig. 3.** PAM distance dependent gap penalties improve sensitivity. For each of 58 real protein sequences, a homolog separated from the sequence by 40 PAM units was embedded in a database of unrelated sequences. A subsequence of 20 amino acids was randomly chosen from the sequence and used to search the database using SSEARCH with a PAM 40 substitution matrix and either default ($q = 10$, $r = 2$) or PAM distance dependent (for PAM40, $q = 21$, $r = 5$) gap penalties. The log change in expectation value ($\log_{10}(E\text{-value}_{\text{default}})$–$\log_{10}(E\text{-value}_{\text{PAM distance dependent}})$ is plotted as a function of the $\log_{10}(E\text{-value}_{\text{default}})$ calculated using default gap penalties.

These gap penalties also improve the alignment of proteins containing short tandemly repeated sequences. We have observed that when aligning proteins with repeated subsequences, the alignments frequently contain spurious gaps. For example, Figure 4 shows a self-alignment of the C-terminal domain (CTD) of mouse RNA polymerase II. This domain contains approximately 50 tandem repeats of a seven amino acid sequence. Using the MDM20 matrix (target distance of 20 PAM units) and gap penalties of $q = 10$ and $r = 2$ produces an alignment with many gaps (vertical or horizontal lines, Figure 4a), where the tandem repeats shift in register by one or two repeat units. Using the gap penalties suggested here, the alignments contained only two or three spurious gaps (Figure 4b). Alignment algorithms are widely used to detect and analyze repeated sequences in protein and DNA (Heringa and Argos, 1993; Benson, 1999; Pellegrini *et al.*, 1999; Matsushima *et al.*, 2000; Andrade *et al.*, 2000); these more stringent gap penalties should improve the detection and characterization of repeated subsequences in protein and DNA.

## DISCUSSION

We have determined effective gap penalties for similarity searching over a range of evolutionary distances from PAM20 to PAM200. We believe the results using real sequences provide the best measure of effective gap penalties, since these results agree well with the simulated data but measure the behavior of real sequences. The agreement of Smith–Waterman results with the those using FASTA indicate that the gap penalties suggested here are also appropriate for use with more widely used heuristic algorithms (Pearson, 2000; Altschul *et al.*, 1997). The gap penalties suggested by these data are higher than those previously used by default in FASTA and those currently used by default in BLAST (Altschul *et al.*, 1997), and should improve performance particularly when short queries are used. The similarities in optimal gap penalties and the independence of the gap extension penalty and PAM distance for real protein homologs are consistent with the observation (Benner *et al.*, 1993) that gap length does not increase significantly with evolutionary distance.

The gap penalties suggested here for PAM matrices can be extrapolated to other substitution matrices, like the BLOSUM series (Henikoff and Henikoff, 1992). Henikoff *et al.* have shown a rough equivalence of a given BLOSUM matrix to a PAM matrix based on relative entropy (Henikoff and Henikoff, 1992). Based on relative entropy, the most effective gap penalties for a BLOSUM80 matrix should be nearly the same as for the PAM120 matrix ($q = 9, r = 3$), and the most effective gap penalties for BLOSUM62 should be similar to those for PAM160 ($q = 5, r = 3$). (In both cases, the penalties implied by the PAM matrices have been scaled to adjust for the 1/2 bit scaling used for the BLOSUM80 and BLOSUM62 matrices.) However, for very distant relationships with full length sequences, effective gap penalties for BLOSUM62 are $q = 7, r = 1$ (Pearson, 1995). Both sets of gap penalties for BLOSUM62 suggest the same penalty ($q+r$) for the first residue in a gap (8), but differ in the extension penalty.

This paper describes the best gap penalties for sequences that have diverged by a target PAM distance. Usually one does not know the target PAM distance, but in some cases, e.g. searching for orthologs between rodents and primates or identification of short-period repeats, a target distance can be estimated. The common practice of using a very 'deep' scoring matrix does not usually prevent one from identifying more closely related homologs (Altschul, 1991; Pearson, 1995), but alignments should improve with scoring matrices and gap penalties targeted to the correct evolutionary distance. These penalties have been incorporated into the current version of the FASTA program.

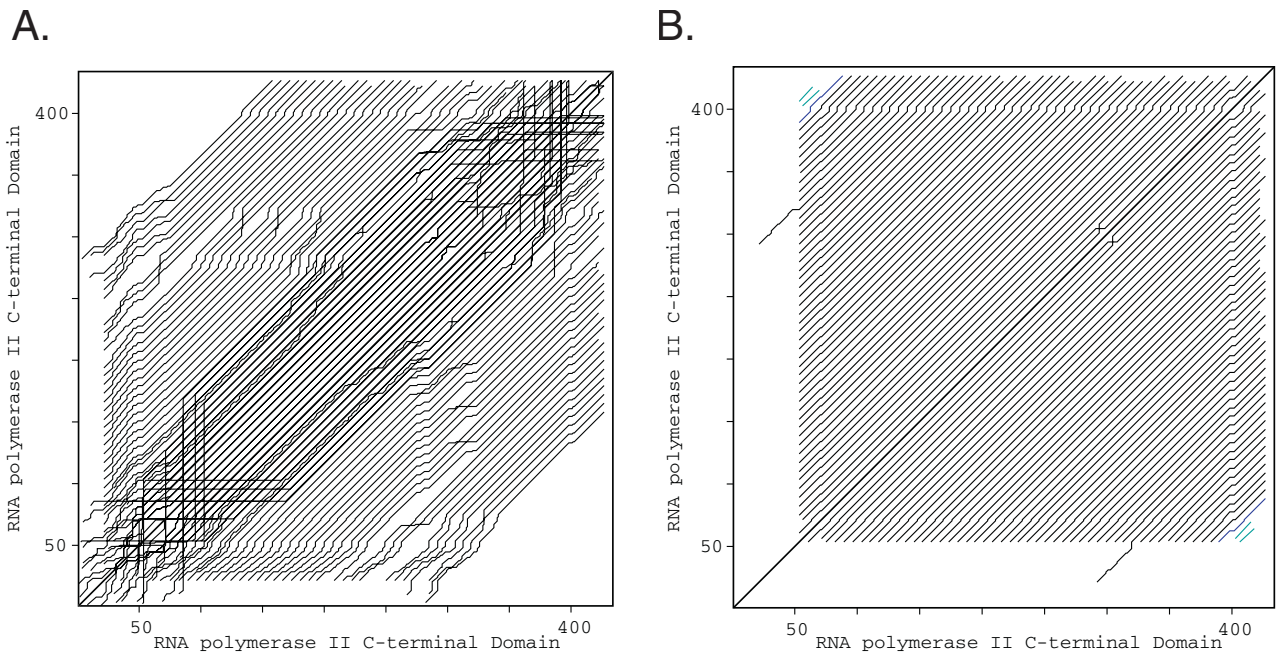


**Fig. 4.** PAM distance dependent gap penalties improve alignments. The C-terminal domain of Mus musculus RNA polymerase II (gi|90464) was aligned to itself with PLALIGN (Pearson, 2000) using an MDM20 substitution matrix and either (A) default gap penalties of $q = 10$ and $r = 2$ or (B) PAM dependent gap penalties of $q = 23$ and $r = 5$.

The gap penalties suggested here maximize sensitivity in database searching, not the quality of the pairwise alignment. While our gap penalties improve the alignment of proteins containing repeated sequences (Figure 4), it is likely that the most effective gap penalties for a given substitution matrix used in database searching differ from the penalties that produce the best alignment between two sequences (Vingron and Waterman, 1994). In the former case, one is attempting to detect homologs by maximizing the difference between the alignment scores with homologous sequences and those with unrelated sequences. In contrast, optimizing an alignment between homologous proteins or protein domains is a global alignment problem. Vingron and Waterman (1994) and others have shown that low gap penalties shift alignments from local to global, so it is not surprising that lower gap penalties can produce more accurate alignments for very distantly related sequences. Vogt *et al.* have determined the gap penalties that produce the best alignment for the PAM matrices, and the optimal gap penalties are in fact lower than the ones reported here (Vogt *et al.*, 1995). Nonetheless, for more closely related sequences with short tandem repeats, low gap penalties can produce less accurate alignments (Figure 4).

This gap penalty formula should not be applied to matrices with target distances greater than 200 PAMs, the longest distance we tested. At PAM 250 the formula recommends $q = 0$ and $r = 5$, which will violate the underlying statistical model by producing global rather than local alignments (Altschul and Gish, 1996). For longer distances, (Pearson, 1995) has determined gap penalties that maximize sensitivity and selectivity. This 'discontinuity' probably reflects a transition from the need for local alignments to reduce the scores of unrelated sequences and the need for global alignments to produce the highest homolog score. The improvement in sensitivity using our PAM-dependent gap penalties will be most dramatic when searching with short sequences, such as those produced in proteomics efforts direct sequencing, or when 'shallower' matrices are used to focus on recent evolutionary events.

## ACKNOWLEDGEMENTS

JTR was supported by an NIH training grant (T32-GM08136). WRP was supported by a grant from the National Library of Medicine (LM04969). The authors thank Ming-qian Huang for the construction of some of the databases used herein.

## REFERENCES

Altschul,S.F. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **219**, 555–565.

Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol.*, **266**, 460–480.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Andrade,M., Ponting,C., Gibson,T. and Bork,P. (2000) Homology-based method for indentification of protein repeats using statistical significance measures. *J. Mol. Biol.*, **298**, 521–537.

Benner,S.A., Cohen,M.A. and Gonnet,G.H. (1993) Empirical and structural models for insertions and deletions in the divergent evolution of proteins. *J. Mol. Biol.*, **229**, 1065–1082.

Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.

Dayhoff,M., Schwartz,R. and Orcutt,B. (1978) A model of evolutionary change in proteins. In Dayhoff,M. (ed.), *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Washington, pp. 345–352.

Fitch,W. and Smith,T. (1983) Optimal sequence alignments. *Proc. Natl Acad. Sci. USA*, **80**, 1382–1386.

Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.

Gu,X. and Zhang,J. (1997) A simple method for estimating the parameter of substitution rate variation among sites. *Mol. Biol. Evol.*, **14**, 1106–1113.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Heringa,J. and Argos,P. (1993) A method to recognize distant repeats in protein sequences. *Proteins*, **17**, 391–441.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.

Matsushima,N., Ohyanagi,T., Tanaka,T. and Kretsinger,R.H. (2000) Super-motifs and evolution of tandem leucine-rich repeats within the small proteoglycans–biglycan, decorin, lumican, fibromodulin, PRELP, keratocan, osteoadherin, epiphycan, and osteoglycin. *Proteins*, **38**, 210–225.

Mott,R. (1992) Maximum-likelihood estimation of the statistical distribution of Smith–Waterman local sequence similarity scores. *Bull. Math. Biol.*, **54**, 59–75.

Pearson,W.R. (1995) Comparison of methods for searching protein sequence databases. *Protein Sci.*, **4**, 1145–1160.

Pearson,W.R. (1998) Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.*, **276**, 71–84.

Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. In Misener,S. and Krawetz,A. (eds), *Methods in Molecular Biology*. Humana Press, Tolowa, NJ, pp. 185–219.

Pearson,W.R. and Wood,T. (2001) Statistical significance in biological sequence comparison. In Balding,D.J., Bishop,M. and Cannings,C. (eds), *Handbook of Statistical Genetics*. Wiley, West Sussex, pp. 39–65.

Pellegrini,M., Marcotte,E.M. and Yeates,T.O. (1999) A fast algorithm for genome-wide analysis of proteins with repeated sequences. *Proteins*, **35**, 440–446.

Vingron,M. and Waterman,M.S. (1994) Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J. Mol. Biol.*, **235**, 1–12.

Vogt,G., Etzold,T. and Argos,P. (1995) An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *J. Mol. Biol.*, **249**, 816–831.

Waterman,M.S., Smith,T. and Beyer,W. (1976) Some biological sequence metrics. *Adv. Math.*, **20**, 367–387.