PROTEINS: Structure, Function, and Bioinformatics 58:329–338 (2005)

A Generalized Affine Gap Model Significantly Improves Protein Sequence Alignment Accuracy

Marcus A. Zachariah,[1] Gavin E. Crooks,[1] Stephen R. Holbrook,[2] and Steven E. Brenner[1,2]*
[1]*Department of Plant and Microbial Biology, University of California, Berkeley, California*
[2]*Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California*

ABSTRACT **Sequence alignment underpins common tasks in molecular biology, including genome annotation, molecular phylogenetics, and homology modeling. Fundamental to sequence alignment is the placement of gaps, which represent character insertions or deletions. We assessed the ability of a generalized affine gap cost model to reliably detect remote protein homology and to produce high-quality alignments. Generalized affine gap alignment with optimal gap parameters performed as well as the traditional affine gap model in remote homology detection. Evaluation of alignment quality showed that the generalized affine model aligns fewer residue pairs than the traditional affine model but achieves significantly higher per-residue accuracy. We conclude that generalized affine gap costs should be used when alignment accuracy carries more importance than aligned sequence length. Proteins 2005;58:329–338.**

Key words: remote homology detection; alignment quality; insertion; deletion; low-similarity region; unaligned

INTRODUCTION

The alignment of biological sequences occupies a central role in modern molecular biology. Fundamental to biological sequence alignment is the incorporation of gaps, which represent insertions or deletions of sequence characters. Needleman and Wunsch[1] introduced the first method of finding the optimal gapped global alignment of two protein sequences. Under the Needleman–Wunsch algorithm, matches between identical or similar characters are assigned positive scores, whereas a penalty is subtracted for each gapped region.

The first gap penalty scheme in common use assigned a cost of bk to each gap, where b is the cost per gapped character and k is the gap length. This length-proportional gap model was supplanted by Smith and Waterman's affine gap cost.[2,3] The affine gap model, which can be implemented with equivalent computational complexity to length-proportional gap costs,[4] charges a penalty of $a + bk$ for each gap, where a is the gap open cost, b is the penalty per gapped character, and k is the gap length. Fitch and Smith[5] argued for the use of affine gaps in place of the simple length-proportional gap model, providing several examples where affine gaps allow for a biologically correct alignment and length-proportional gaps do not. The superiority of affine over length-proportional gap costs is reflected by the nearly ubiquitous use of affine gaps by pairwise methods (e.g., BLAST,[6] FASTA,[7] and SSEARCH[7]). However, although easy to implement and fast to calculate, traditional affine gaps almost certainly do not model the evolution of insertions and deletions very closely.[8–12] For this reason, new gap models have continued to be developed.[8–14] Qian et al.[12] provided evidence that a quadruple affine gap penalty would better model the distribution of gap lengths in structural alignments. Altschul[11] introduced a generalized affine gap model that allows for the inclusion of unaligned regions within larger alignments. We focus on this model in the present work.

Observations of protein structure motivated the development of generalized affine gap costs. Distantly related proteins often share sequence and structure similarity in functionally important regions but have diverged elsewhere (see Fig 1). Frequently, these divergent regions lack any meaningful alignment, and attempting to optimally align them detracts from the overall quality of a sequence alignment. Generalized affine gaps aim to leave dissimilar regions in proteins unaligned in order to better align regions that are more closely related. In addition to this gap model, other methods of exploiting low-similarity segments of protein sequence alignment have been introduced.[17,18]

Like traditional affine gap costs, the generalized affine model charges a fixed cost for the existence of a gap (which we call a), as well as a cost per gapped residue (which we call b). Unaligned residue pairs may be included in the gap and charged an additional penalty (c). The cost for a gap involving k_1 residues in one sequence and k_2 in the other, with $k_1 \geq k_2$, is $a + b(k_1 - k_2) + ck_2$. Unlike many other recently proposed gap models,[8–10,13] generalized affine

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html

Grant sponsor: U.S. Department of Energy; Grant number: DE-AC03-76SF00098. Grant sponsor: National Institutes of Health; Grant number: 1 K22 HG00056. Grant sponsor: Sloan/DOE postdoctoral fellowship in computational molecular biology (to G. E. Crooks). Grant sponsor: Searle; Grant number: 1-L-110 (to S. E. Brenner).

*Correspondence to: Steven E. Brenner, Department of Plant and Microbial Biology, 461A Koshland Hall, Ste. 3102, Berkeley, CA 94720-3102. E-mail: brenner@compbio.berkeley.edu

Received 13 March 2004; Accepted 14 July 2004

Published online 23 November 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20299

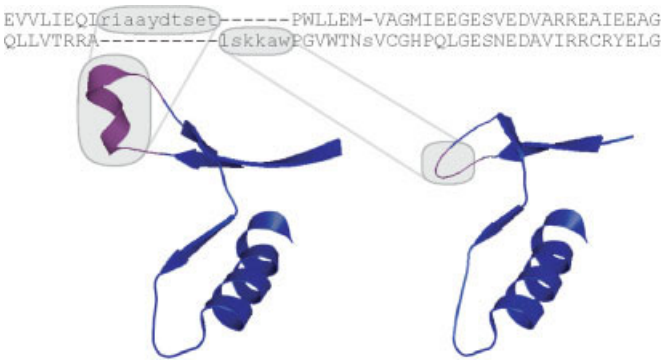

Fig. 1. Movitation for generalized affine gap costs. Shown here are fragments of 2 nudix proteins [**left**, PDB ID: 1G0S[15]; **right**, PDB ID: 1HZT[16]] that are homologous but possess no recognizable similarity in the loop region. The lack of similarity in such regions often impedes accurate and meaningful alignment.

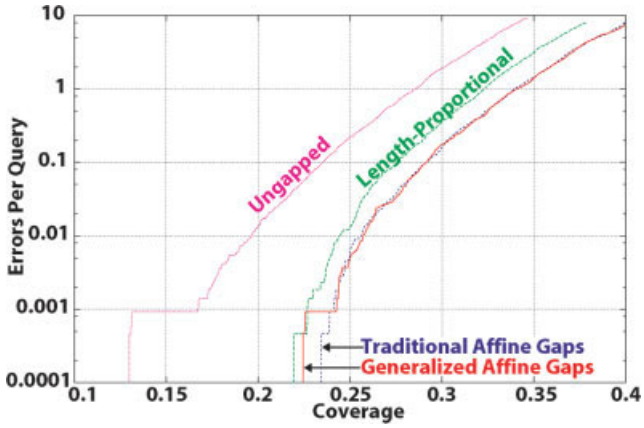

Fig. 4. Coverage versus error of 4 gap cost models. After selecting the optimal amino acid substitution matrix and gap parameters for each model (see Table II), we compared their performance on the test set. Remote homology detection coverage is shown at various error rates.

gaps may be incorporated into global Needleman–Wunsch[1] or local Smith–Waterman[2] alignment without increasing the space or time complexity. See Altschul[11] for a more detailed description.

Altschul evaluated the performance of the generalized affine gap model in remote homology detection and alignment quality. Testing distant homolog identification, he selected parameters for the traditional and generalized affine gap models that resulted in the two models giving identical distributions of scores to unrelated sequences. Notably, when 222 evolutionarily related sequence pairs were aligned with these gap parameters, generalized affine gaps assigned average statistical scores 50% higher than those assigned by traditional affine gaps. This suggests that generalized affine gaps may offer improved homology detection over the traditional affine model. Assessing alignment quality in a test of 26 sequence pairs, Altschul demonstrated that generalized affine gap alignments conform better to PSI-BLAST[19] reference alignments than do traditional affine gap alignments. This finding hints at a possible improvement in alignment accuracy.

Altschul's results argue for the inclusion of generalized affine gaps in methods that employ pairwise sequence comparison, such as BLAST and PSI-BLAST. However, the more recent study of Schaffer et al.,[20] examining 653,123 profile–sequence comparisons, reported that generalized affine gaps bring no improvement in homology detection ability to PSI-BLAST. The lack of consensus and the use of relatively small test sets left the suitability of this theoretically attractive gap model in doubt. Additionally, neither study attempted to find optimal parameters for the generalized model, and neither employed a sequence-independent standard for benchmarking. Sequence-dependent standards create a "chicken and egg problem,"[21] in which newer methods may be penalized for correctly identifying results missed by older methods.

In this study, we rigorously evaluate the homology detection ability of pairwise local alignment using generalized affine gap costs. The SCOP database,[22,23] a set of proteins whose evolutionary relationships have been inferred through structural analysis, forms the basis of our evaluation (see Methods section). SCOP provides a sequence-independent standard for benchmarking, and it allows us to make 4,761,124 sequence comparisons: 21,446-fold more than employed by Altschul and 7-fold more than employed by Schaffer et al. After determining the optimal gap parameters for each model, we show that generalized affine gaps offer a statistically insignificant improvement in homology detection over traditional affine gaps. We also assess, using 1000 pairwise structural alignments from the FSSP database,[24] the quality of alignments produced by the traditional and generalized affine gap models. The FSSP test set contains 38-fold more alignments than used in Altschul's study of alignment quality. Moreover, its structural alignments represent a sequence-independent benchmark. Our results indicate that the generalized model is more conservative than the traditional model in aligning residues, and consequently achieves significantly higher alignment accuracy. We conclude by suggesting applications where generalized affine gap costs should be used.

METHODS

Remote Homology Detection

To assess each gap model's ability to reliably detect evolutionarily distant protein homology, we performed the following analysis. First, a set of proteins whose evolutionary interrelations are known was assembled from the SCOP database (version 1.61),[22,23] which provides a hierarchical classification of the structural domains of all solved protein structures. SCOP categorizes protein domains at the level of class, fold, superfamily, and family. If two domains belong to different classes or folds, they may safely be considered unrelated.[25] When of the same superfamily or family, proteins are considered homologous.[25] We treat the evolutionary relationship of domains classified in the same fold but different superfamily as undetermined, and do not consider them in our benchmarking.

To focus our study on distant homologs, we used only protein domain sequences in SCOP not more than 40%

TABLE I. Evaluated Gap Parameters and Amino Acid Substitution Matrices

Gap Model	Amino Acid Substitution Matrix	Gap Open Penalty (a)	Penalty per Gapped Residue (b)	Penalty per Unaligned Residue Pair (c)
Ungapped $g(k) = \infty$	BLOCKS 13+ BLOSUM65 VTML 190 VTML 210 VTML 240	∞	∞	∞
Length-proportional $g(k) = bk$	BLOCKS 13+ BLOSUM65 VTML 190 VTML 210 VTML 240	0	1, 4, 7. . .67	∞
Traditional affine $g(k) = a + bk$	BLOCKS 13+ BLOSUM65 VTML 190 VTML 210 VTML 240	50, 60, 70, 80, 90, 100, 110, 120, 130, 140	4, 7, 10, 13	∞
Generalized affine $g(k_1, k_2) = a + b(k_1 - k_2) + ck_2$	BLOCKS 13 + BLOSUM65 VTML 190 VTML 210 VTML 240	50, 60, 70, 80, 90, 100, 110, 120, 130, 140	4, 7, 10, 13	0, 3, 6, 9, 12, 15, 18, 21, 24, 27, 30

The given gap parameters and substitution matrices were assessed for remote homology detection within the training set under the gap models indicated. In order to explore small integer gap penalties, we used 1/20 bit scaling for all parameters and matrices.

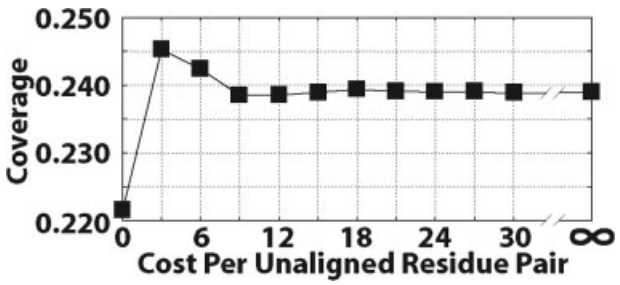

Fig. 2. Coverage at increasing penalty per unaligned residue pair. Remote homology detection coverage (at 0.01 errors per query) peaks at penalty per unaligned residue pair = 3, and decreases as the penalty is raised. When the unalign penalty is infinite, the generalized affine gap model reduces to traditional affine gaps. Shown here are data from local alignment on the training set with the VTML190 substitution matrix (1/20 bits), gap open penalty = 80, and penalty per gapped residue = 10.

identical to each other. The ASTRAL compendium[26–28] conveniently provides such a set of SCOP protein domains. The 40%-filtered set was divided into training (2592 sequences) and test (2182 sequences) subsets, where each subset contains all sequences of every other fold, and there are no sequences in the intersection of the two subsets. Sequences are available at http://compbio.berkeley.edu.

We evaluated gap models by using them to perform local alignments of all pairs of sequences in a given data set using several different substitution matrices. These were BLOSUM65, reparameterized from the BLOCKS 13+ database[29,30] [which is the most effective BLOSUM matrix on this data set (Price, Crooks, Green, and Brenner, unpublished)], and several VTML matrices[31,32] representing different divergence times. VTML is essentially a reparameterized Dayhoff PAM matrix and, as such, unit time corresponds to 1% point accepted mutation (PAM). For each sequence an ordered list of putative homologs was generated by ranking alignments by *e*-value. These significance scores were calculated using the method of Bailey and Gribskov.[33] An *e*-value cutoff can be chosen to dictate which members of the list are (correctly or incorrectly) considered to be homologs. As in Brenner et al.[21] and Green and Brenner,[25] we measured the coverage and errors per query for every cutoff. Coverage is the fraction of true homologous relationships (true positives) found by the method to be evaluated. Errors per query, a measure of the false-positive rate, indicate the number of sequence pairs incorrectly found to be homologous by the alignment method divided by the number of sequences in the data set.

The number of relationships within a given superfamily grows quadratically with superfamily size. Therefore, any representational biases present within the database are exacerbated, and large families dominate the overall results.[25] To compensate for this, we reweighted each correct pairwise relation by $1/(n-1)$, where n is the number of sequences and $(n-1)$ is the number of true homologs per sequence in a superfamily. Thus, each superfamily was weighted in linear proportion to its size.

The statistical significance of our results were estimated with Bayesian bootstrap resampling[25,34] (Price, Crooks, Green, and Brenner, unpublished). In brief, we generate 500 replicas of the original sequence data set. In a traditional

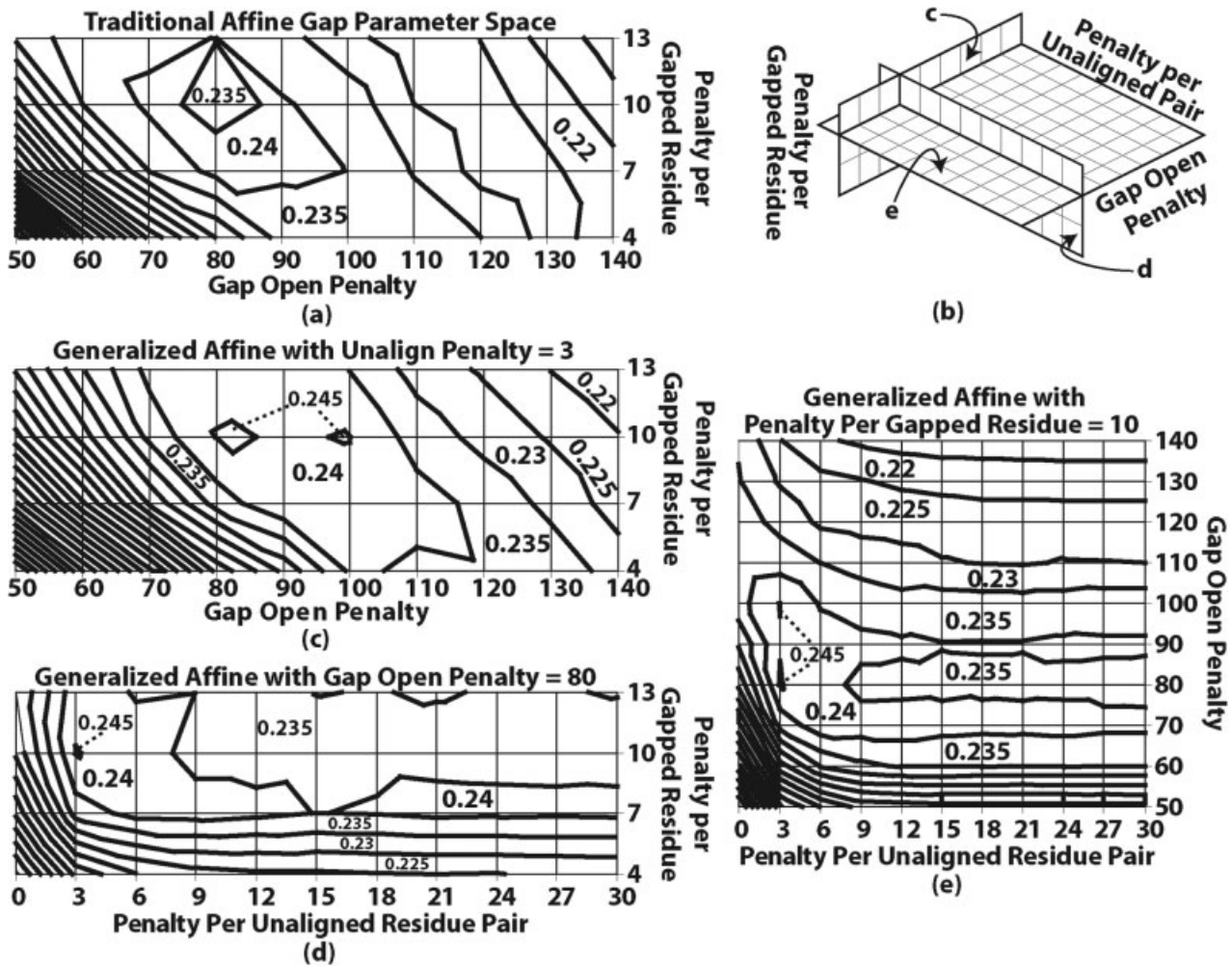

Fig. 3. Comparison of traditional affine and generalized affine gap parameter space. (**a**) We determined coverage on the training set at 0.01 errors per query for a range of traditional affine gap parameters. The peak coverage of 24.15% occurs at gap open penalty = 80 and penalty per gapped residue = 7. (**b**) To identify the optimal parameters for generalized affine gaps, we searched over a three-dimensional space. The spatial relationship between the two-dimensional planes given in (**c**), (**d**), and (**e**) is depicted. (**c–e**) Shown here are orthogonal slices through the generalized affine gap parameter space. Slices intersect at the coverage peak (25.52%), which is found at gap open penalty = 80, penalty per gapped residue = 10, and penalty per unaligned pair = 3. The penalty per unaligned residue pair is fixed at 3 in (**c**); the gap open penalty in (**d**) is 80, and the penalty per gapped residue in (**e**) is 10. All data are from local alignment using the VTML190 substitution matrix.

bootstrap with resampling, each sequence is included in a replica 0, 1, 2, or more times. However, for our application, the zero-occupancies introduce significant bias into the results. Therefore, rather than giving sequence-discrete weights, within each replica, each sequence is assigned a weight drawn from a Dirichlet distribution. Consequentially, each data replica generates a different answer, and the replica ensemble provides a Bayesian posterior estimate of the correct result. Using the paired t test, we compared the ability of each model to reliably identify homologous relationships within the 500 resampled data sets. At a given error rate, we report the observed difference in coverage and a confidence interval that indicates the significance of the result.

Alignment Quality

We assessed the quality of pairwise alignments gapped under either scheme by comparing them against trusted structural alignments derived from the FSSP database.[24] Our first set of trusted alignments, which we call ES1, is a subset of the compilation used by Edgar and Sjölander,[35] whose data set contains alignments between pairs in FSSP meeting the following criteria: pairwise identity $\leq$ 30%, Dali[36] Z score $\geq$ 15, root-mean-square deviation (RMSD) $\leq$ 2.5 Å, and agreement between the combinatorial extension (CE)[37] and Dali structural aligners along at least 50 aligned residue pairs. Alignments were filtered so that no two sequences aligned to a common third sequence had greater than 30% identity. Of the resulting 588 alignments, we randomly chose 500 to include in our analysis.

The ES1 set of aligned sequences contains pairs of high structural similarity. To include more structurally diverged proteins, we employed a second data set (ES2), also derived from FSSP and assembled by Edgar and Sjölander.[38] They selected sequence pairs that met the following criteria: pairwise identity $\leq$ 30%, Dali Z score $\geq$ 8 and $\leq$

TABLE II. Optimal Gap Parameters and Amino Acid Substitution Matrices

Ungapped Alignment

Substitution Matrix	Optimum Gap Parameters	Coverage
VTML190	—	0.173
VTML240	—	0.166
BLOCKS 13 + BLOSUM65	—	0.158
VTML210	—	0.154

Length-Proportional Gap Model

Substitution matrix	Optimum Gap Parameters (per gapped residue)	Coverage
VTML190	58	0.225
VTML240	52	0.222
VTML210	61	0.214
BLOCKS 13 + BLOSUM65	58	0.208

Traditional Affine Gap Model

Substitution Matrix	Optimum Gap Parameters (open/per gapped residue)	Coverage
VTML190	80/7	0.241
VTML240	80/7	0.239
VTML210	90/10	0.234
BLOCKS 13 + BLOSUM65	90/10	0.225

Generalized Affine Gap Model

Substitution Matrix	Optimum Gap Parameters (open/per gapped residue/per unaligned pair)	Coverage
VTML190	80/10/3	0.245
VTML240	80/10/3	0.242
VTML210	90/10/6	0.236
BLOCKS 13+ BLOSUM65	110/7/3	0.227

For each gap model and substitution matrix, the gap parameters achieving the highest coverage of remote homologs within the training set at 0.01 errors per query are presented along with the relevant coverage values.

12, RMSD $\leq$ 3.5 Å, and alignment length $\geq$ 50. Because many of the resulting pairs contained sequences of questionable evolutionary relatedness, they additionally required that aligned sequences be homologous according to the SCOP test. Sequences were filtered in the same manner as in the first set, and 500 of the resulting alignments were selected at random to use in benchmarking.

We used two scores to measure the quality of a given sequence alignment with respect to a gold standard reference alignment. The modeler's score,[39,40] also referred to as SP,[35,41] represents the number of correctly aligned residue pairs divided by the length of the reference alignment. This measure does not penalize overalignment. We also calculated the developer's score,[39,40] called PS by others,[35,41] as the number of correctly aligned residue pairs divided by the length of the alignment being tested. This measure does not penalize underalignment, and neither score discriminates between slightly offset and completely incorrect alignments. In the case of perfect agreement between the test and reference alignments, both scoring methods produce their maximum value: 1. The developer's and modeler's scores possess a minimum value of 0 for alignments in which no residue pairs are correctly aligned.

RESULTS AND DISCUSSION

Remote Homology Detection

To conduct an unbiased evaluation, we divided our data set into test and training subsets (see Methods section). We used the training set to evaluate the remote homology detection ability of a range of gap parameters and amino acid substitution matrices (in 1/20 bit scaling)[30–32] to be used in conjunction with the gap models (see Table I). Each sequence in the training set was compared to every other, and putative homologs were ranked by *e*-value (see Methods section). Following previous studies,[21,25] we report the proportion of correctly identified homologs (coverage) at an error rate of 0.01 errors per query. Coverages have been linearly normalized to correct for representational biases inherent in SCOP (see Methods section and Green and Brenner[25]). As the traditional affine gap model is a special case of the generalized model with unalign penalty = ∞, we examined remote homology detection as the penalty per unaligned residue pair is increased from 0 to ∞ (see Fig. 2). Coverage of remote homologs at the 1% error rate rises as we move away from unalign penalty = 0, peaks at unalign penalty = 3, and slowly decreases as we move toward unalign penalty = ∞ (equivalent to traditional affine gaps). This trend was seen with all 4 substitution matrices.

Analysis of all training set results (included as Supplementary Data) indicates that among the 4 tested substitution matrices (see Table I), VTML190[31,32] best detects remote homology, regardless of the gap model. For traditional and generalized affine gap costs, gap open penalties ranging from 80 to 100, and penalties per gapped residue of 7 or 10 consistently produce the best results (see Fig. 3). The traditional affine gap model reaches its peak coverage of 24.15% with the VTML190 matrix, gap open cost = 80, and penalty per gapped residue = 7. The generalized affine gap model reaches a higher peak coverage of 24.52% with the same matrix, same gap open cost, penalty per gapped residue = 10, and penalty per unaligned residue pair = 3. Table II presents the optimum parameters and substitution matrix for each gap model. Finally, we note that when the penalty per unaligned residue pair is twice the penalty per gapped residue, the generalized affine gap model reduces to the commonly used traditional affine gap implementation that allows a gap in one sequence immediately adjacent to a gap in the other. While the traditional affine model, as strictly defined, does not allow contiguous gapped regions in opposite sequences,[42] we detected no difference in remote homology detection between the commonly used

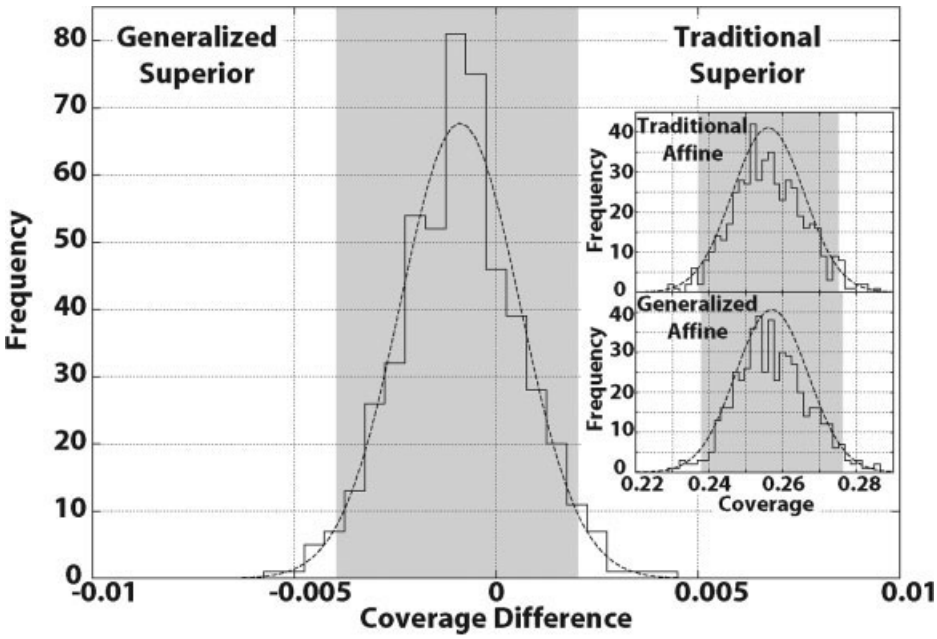

Fig. 5. Bootstrap significance tests. Our test data set was resampled under the Bayesian bootstrap (see Methods section) to give 500 new data sets, and we searched for homologous pairs within each using the traditional and generalized affine gap models. The large histogram gives the distribution of differences in coverage at 0.01 errors per query for a given data set (traditional affine coverage and generalized affine coverage). The large dotted curve is a Gaussian distribution fitted to the coverage differences, and the shaded area contains the 95% confidence interval. Because zero lies within the confidence interval, we deem the difference in coverage statistically insignificant. The inset figures include histograms, fitted Gaussian curves, and confidence intervals for the 2 gap models' individual distributions of coverage on the resampled test data sets.

and strictly defined implementations of traditional affine gap penalties (see Supplementary Data).

Using the gap parameters and amino acid substitution matrices that performed best on the training set, we evaluated the ability of all 4 gap cost models to detect distant evolutionary relationships within the test set (see Fig. 4). Following the progression of gap models from less sophisticated to more sophisticated, we see diminishing improvements in performance. Local alignment with length-proportional gap costs offers substantial improvement over ungapped local alignment, increasing coverage at the 1% error rate (0.01 errors per query), from 19.6% to 24.4%. Traditional affine gaps achieve a smaller gain, raising coverage at the same error rate to 25.5%. Generalized affine gaps bring a miniscule elevation in coverage to 25.7%. These results can be compared to the coverage difference between BLAST (21.7%) and FASTA (23.8%). Based on the trend of decreased performance gain with increased gap model sophistication, we predict that the more complex gap models presented elsewhere will not substantially improve remote homolog detection.

To assess the statistical significance of the small observed difference in homology detection between traditional and generalized affine gaps, we employed the Bayesian bootstrap (Price, Crooks, Green, and Brenner, unpublished). Under the bootstrap procedure, we resampled our test data set 500 times. For every resampled data set, we evaluated the difference in coverage between the traditional and generalized models at the 1% error rate. A histogram of all 500 coverage differences is shown in Figure 5, with the individual coverages given in the two inset histograms. The coverage differences were fitted to a normal curve with a mean of -8.90×10^{-4} and a standard deviation of 1.47×10^{-3}. Because 0 (no difference) lies within the 95% confidence interval of coverage difference ($[-3.78 \times 10^{-3}, 2.00 \times 10^{-3}]$), the difference in performance between the 2 gap models is not statistically significant (P value = 0.545). As a positive control, we confirmed that our bootstrap procedure detects a statistically significant performance difference between length-proportional and traditional affine gap costs (data not shown).

Alignment Quality

We measured the quality of protein sequence alignments gapped under the traditional and generalized affine models as compared to reference structural alignments compiled by Edgar and Sjölander[35,38] from the FSSP database.[24] The ES1 data set[35] contains 500 pairwise structural alignments between proteins of less than 30% sequence identity but of high structural similarity. The ES2 data set[38] differs from ES1 in that its pairs of proteins have diverged structurally and thus contain more insertions or deletions (see Methods section). Using gap parameters and substitution matrices optimized for homology detection, we generated local alignments with traditional and generalized affine gaps for the pairs of proteins in both data sets. Next, we scored these alignments against the

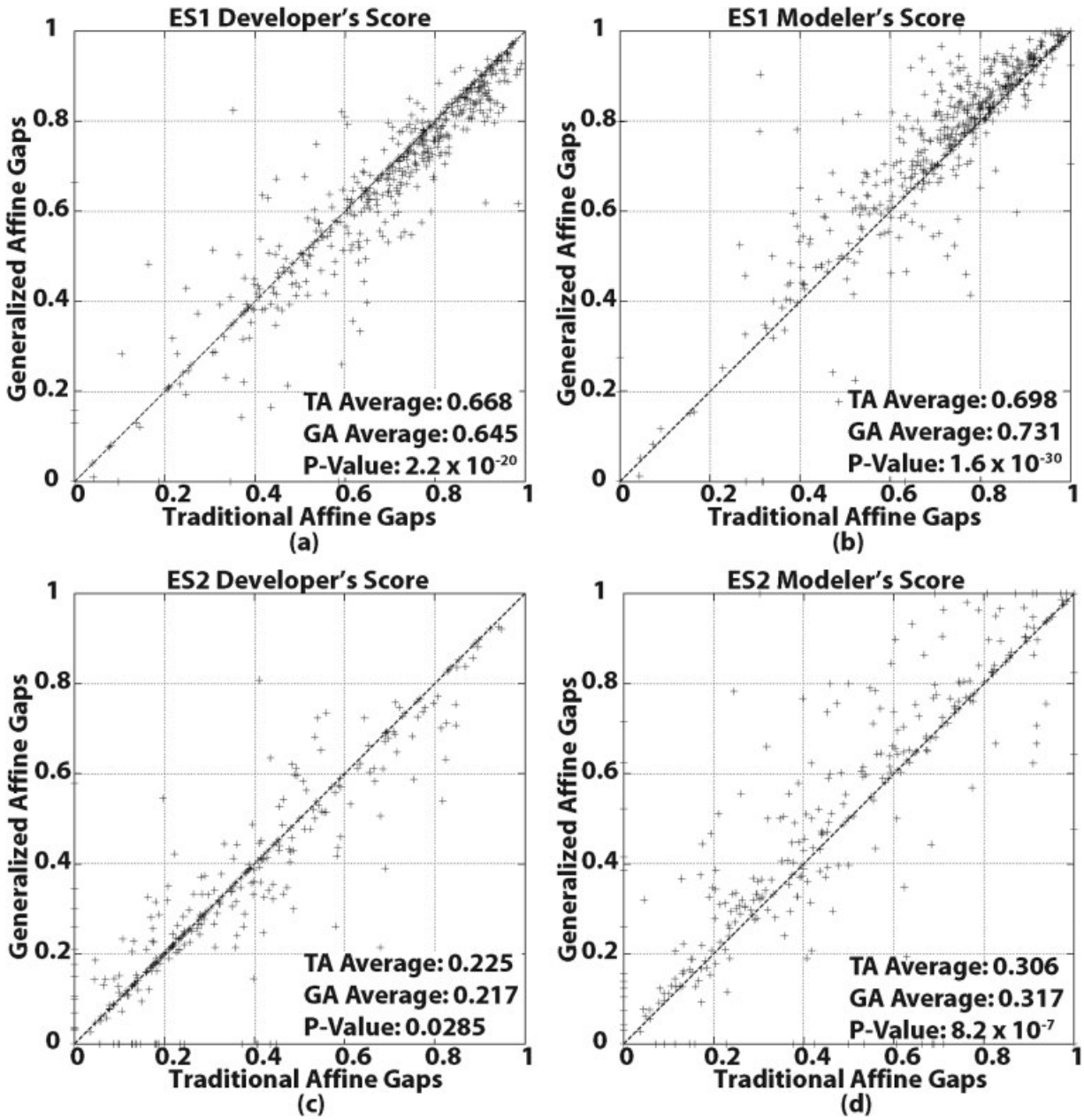

Fig. 6. Alignment quality scores. Developer's (**a**) and (**c**) and modeler's (**b**) and (**d**) scores are given for pairwise alignments of sequences in the ES1 and ES2 data sets. The vertical axis indicates the scores of the generalized affine gaps alignment, and the horizontal axis, the score of the traditional affine gaps alignment. Inset in each plot are the average score of the generalized affine gaps alignments (GAs), the average score of the traditional affine gaps alignments (TAs), and the *P* value of a Wilcoxon signed-rank test comparing the scores of the two gap models.

FSSP reference structural alignments using two measures of alignment quality. The developer's score[39,40] measures the fraction of the reference structural alignment that is correctly aligned and included in the sequence alignment being tested. This score is calculated as (number of aligned residue pairs in the test alignment that are correctly aligned with respect to the reference alignment)/(number of aligned residue pairs in the reference alignment). A second measure of quality, the modeler's score,[39,40] indicates the fraction of residue pairs in the test alignment that are aligned correctly with respect to the reference alignment. It is defined as (number of residue pairs in the test alignment that are correctly aligned with respect to the reference alignment)/(number of aligned residue pairs in the test alignment).

On the ES1 data set, traditional affine gaps outperformed generalized affine gaps when measured by the developer's score [see Fig. 6(a)]. Traditional affine gaps scored on average 0.6677, whereas generalized affine gaps averaged 0.6452. This difference was statistically significant, as shown by the two-tailed Wilcoxon signed-rank test[43] (P value $= 2.2 \times 10^{-20}$). When alignments were

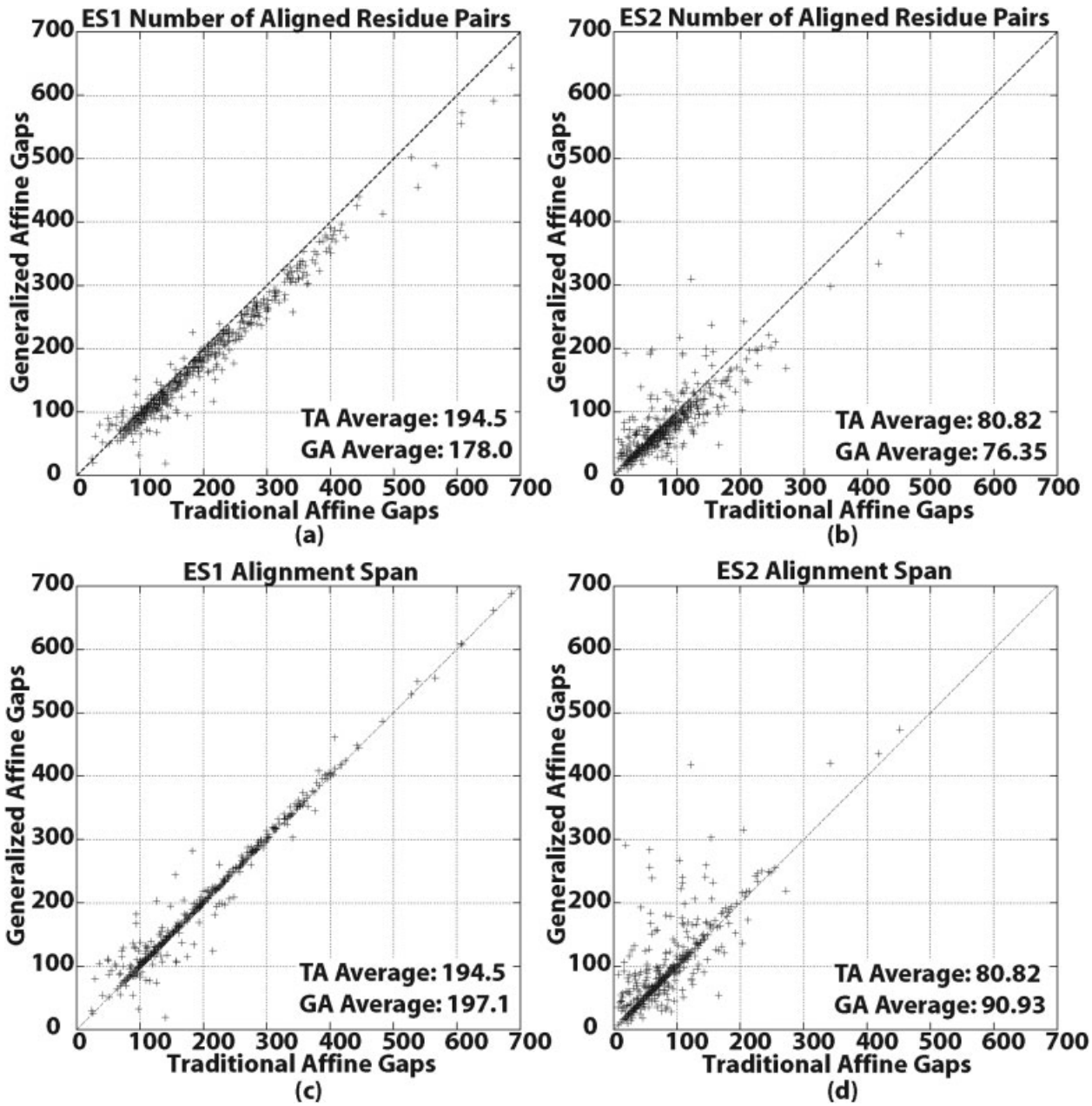

Fig. 7. Alignment lengths. Number of aligned residue pairs is presented in (**a**) and (**b**), with generalized affine gaps (GAs) on the vertical axis and traditional affine gaps (TAs) on the horizontal axis. Total number of residue pairs included in an alignment (both unaligned and aligned) is given in (**c**) and (**d**). Averages are inset in the plots.

measured by the modeler's score, generalized affine gaps outperformed traditional affine gaps [see Fig. 6(b)]. The average scores over all sequence pairs in the ES1 data set were 0.6978 and 0.7311 for the traditional and generalized models, respectively. Again, the two-tailed Wilcoxon test demonstrated the statistical significance of the score difference (P value = 1.6×10^{-30}).

We repeated the analysis on the ES2 data set and found similar results [see Figs. 6(c and d)]. Traditional affine gaps outperformed generalized affine gaps on the developer's score (0.2245 vs 0.2173 average), whereas generalized affine gaps earned the higher modeler's score (0.3056 vs 0.3171 average). The P value for the developer's score comparison was 0.0285; the modeler's score was 8.2×10^{-7}. These results indicate that the generalized affine gap model aligns fewer residue pairs but is more precise in accurately matching the pairs it does align.

Examination of the lengths of alignments confirmed that the generalized affine model aligns fewer residue pairs than the traditional affine model aligns [see Fig. 7(a and b)]. This difference did not stem from the generalized alignments covering a shorter portion of the protein se-

quences being aligned; generalized affine gaps covered at least as large a portion of the concerned sequences as traditional affine gaps covered [see Fig. 7(c and d)]. The difference in number of aligned residue pairs is a result of generalized affine gaps' unique ability to include unaligned residues within larger alignments.

CONCLUSIONS

This article evaluates the remote homology detection ability and alignment quality of generalized affine gap costs for protein sequence alignment. We identify the amino acid substitution matrix and gap parameters optimal for pairwise remote homolog detection with generalized affine gaps, and show that generalized affine gaps offer a statistically insignificant performance advantage over the currently used traditional affine model. This finding agrees with recent work suggesting that increasingly sophisticated methods of pairwise sequence comparison offer little or no improvement in remote homology detection ability over established methods (Price, Crooks, Green, and Brenner, unpublished results).

Alignments produced by the generalized affine gap model include fewer aligned residue pairs but attain significantly higher per-residue accuracy than traditional affine gaps alignments. It is notable that the generalized model uses less of the available sequence information, aligning on average 7.6% fewer residue pairs, but reliably identifies distant evolutionary relationships as well as traditional affine gap costs. This suggests that generalized affine gap costs align the residues most important for determining evolutionary relatedness. It also suggests that the improved alignment of related regions compensates for the loss of information due to unaligned residue pairs. We advocate the use of generalized affine gaps costs for protein sequence alignment where alignment accuracy carries more importance than number of aligned residues.

ACKNOWLEDGMENTS

Our thanks to R. E. Green and E. E. Hill for valuable discussion, and to R. C. Edgar for his comments and for providing the ES1 and ES2 data sets.

REFERENCES

1. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
2. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
3. Altschul SF, Erickson BW. Optimal sequence alignment using affine gap costs. Bull Math Biol 1986;48:603–616.
4. Gotoh O. An improved algorithm for matching biological sequences. J Mol Biol 1982;162:705–708.
5. Fitch WM, Smith TF. Optimal sequence alignments. Proc Natl Acad Sci USA 1983;80:1382–1386.
6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
7. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988;85:2444–2448.
8. Benner SA, Cohen MA, Gonnet GH. Empirical and structural models for insertions and deletions in the divergent evolution of proteins. J Mol Biol 1993;229:1065–1082.
9. Gonnet GH, Cohen MA, Benner SA, Mott R, Metzler D, Wrabl JO, Grishin NV. Exhaustive matching of the entire protein sequence database. Science 1992;256:1443–1445.
10. Mott R. Local sequence alignments with monotonic gap penalties. Bioinformatics 1999;15:455–462.
11. Altschul SF. Generalized affine gap costs for protein sequence alignment. Proteins 1998;32:88–96.
12. Qian B, Goldstein RA, Bailey TL, Gribskov M. Distribution of Indel lengths. Proteins 2001;45:102–104.
13. Metzler D. Statistical alignment based on fragment insertion and deletion models. Bioinformatics 2003;19:490–499.
14. Wrabl JO, Grishin NV. Gaps in structurally similar proteins: towards improvement of multiple sequence alignment. Proteins 2004;54:71–87.
15. Gabelli SB, Bianchet MA, Bessman MJ, Amzel LM. The structure of ADP-ribose pyrophosphatase reveals the structural basis for the versatility of the Nudix family. Nat Struct Biol 2001;8:467–472.
16. Durbecq V, Sainz G, Oudjama Y, Clantin B, Bompard-Gilles C, Tricot C, Caillet J, Stalon V, Droogmans L, Villeret V. Crystal structure of isopentenyl diphosphate:dimethylallyl diphosphate isomerase. EMBO J 2001;20:1530–1537.
17. Edgar RC, Sjölander K. SATCHMO: sequence alignment and tree construction using hidden Markov models. Bioinformatics 2003;19:1404–1411.
18. Sunyaev SR, Bogopolsky GA, Oleynikova NV, Vlasov PK, Finkelstein AV, Roytberg MA. From analysis of protein structural alignments toward a novel approach to align protein sequences. Proteins 2004;54:569–582.
19. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
20. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;29:2994–3005.
21. Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci USA 1998;95:6073–6078.
22. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
23. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data [Database issue]. Nucleic Acids Res 2004;32:D226–D229.
24. Holm L, Sander C. Mapping the protein universe. Science 1996;273:595–603.
25. Green RE, Brenner SE. Bootstrapping and normalization for enhanced evaluations of pairwise sequence comparison. Proc IEEE 2002;90:1834–1847.
26. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 2000;28:254–256.
27. Chandonia JM, Walker NS, Lo Conte L, Kohel P, Levitt M, Brenner SE. ASTRAL compendium enhancements. Nucleic Acids Res 2002;30:260–263.
28. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL compendium in 2004 [Database issue]. Nucleic Acids Res 2004;32:D189–D192.
29. Henikoff S, Henikoff JG. Amino-acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
30. Henikoff JG, Greene EA, Pietrokovski S, Henikoff S. Increased coverage of protein families with the blocks database servers. Nucleic Acids Res 2000;28:228–230.
31. Muller T, Vingron M. Modeling amino acid replacement. J Comput Biol 2000;7:761–776.
32. Muller T, Spang R, Vingron M. Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. Mol Biol Evol 2002;19:8–13.
33. Bailey TL, Gribskov M. Estimating and evaluating the statistics of gapped local-alignment scores. J Comput Biol 2002;9:575–593.
34. Rubin DB. The Bayesian bootstrap. Ann Stat 1981;9:130–134.
35. Edgar RC, Sjölander K. A comparison of scoring functions for protein sequence profile alignment. Bioinformatics 2004;20:1301–1308.
36. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.

37. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
38. Edgar RC, Sjölander K. Coach: profile–profile alignment of protein families using hidden Markov models. Bioinformatics 2004;20: 1309–1318.
39. Sauder JM, Arthur JW, Dunbrack RL Jr. Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins 2000;40:6–22.
40. Kahsay RY, Wang G, Dongre N, Gao G, Dunbrack RL Jr. CASA: a server for the critical assessment of protein sequence alignment accuracy. Bioinformatics 2002;18:496–497.
41. Thompson JD, Higgins DB, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
42. Durbin R, Eddy SR, Krogh A, Mitchison GJ. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge, UK: Cambridge University Press; 1998.
43. Gotoh O. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. J Mol Biol 1996;264:823–838.