Bioinformatics Advance Access published August 31, 2013

Adjusting Scoring Matrices to Correct Overextended Alignments

Lauren J. Mills[1], and William R. Pearson[2*]

[1]University of Virginia, Department of Molecular, Cell and Developmental Biology, Box 800793, Charlottesville, VA 22908
[2]University of Virginia, Department of Biochemistry and Molecular Genetics, Box 800733, Charlottesville, VA 22908

Associate Editor: Dr. John Hancock

ABSTRACT

Motivation: Sequence similarity searches performed with BLAST, SSEARCH, and FASTA achieve high sensitivity by using scoring matrices (e.g. BLOSUM62) that target low identity (<33%) alignments. While such scoring matrices can effectively identify distant homologs, they can also produce local alignments that extend beyond the homologous regions.

Results: We measured local alignment start/stop boundary accuracy using a set of queries where the correct alignment boundaries were known, and found that 7% of BLASTP and 8% of SSEARCH alignment boundaries were over-extended. Over-extended alignments include non-homologous sequences; they occur most frequently between sequences that are more closely related (>33% identity). Adjusting the scoring matrix to reflect the identity of the homologous sequence can correct higher-identity over-extended alignment boundaries. In addition, the scoring matrix that produced a correct alignment could be reliably predicted based on the sequence identity seen in the original BLOSUM62 alignment. Realigning with the predicted scoring matrix corrected 37% of all over-extended alignments, resulting in more correct alignments than using BLOSUM62 alone.

Availability: RPD2 sequences and the FASTA software are available from http://faculty.virginia.edu/wrpearson/fasta

Contact: wrp@virginia.edu

1 INTRODUCTION

Sequence similarity search algorithms are used to identify evolutionary homologs and to generate hypotheses for the function of unknown proteins. These algorithms assign homology between sequences achieving statistically significant similarity scores with high fidelity, even between highly divergent sequences sharing low similarity (Pearson, 1995; Brenner *et al.*, 1998; Pearson and Sierk, 2005). However, the same methodology that provides for the sensitive identification of homology at low identity can also lead to alignments that include non-homologous sequence adjacent to, or between, higher identity homologous sequences (Gonzalez and Pearson, 2010a).

Homologous over-extension was first identified as a source of error during iterative similarity searches (Gonzalez and Pearson, 2010a). Over-extension occurs when alignments extend past the boundaries of the homologous region in the library, query, or both sequences, leading to the inclusion of non-homologous sequence in an alignment (Fig. 1). The inclusion of non-homologous sequence has been identified in alignments between highly identical DNA sequences (Chao *et al.*, 1993) and has been termed the "mosaic effect" (Arslan *et al.*, 2001).

Over-extension occurs because local sequence alignment boundaries depend on the scoring matrix. The popular BLASTP (Altschul *et al.*, 1997) tool, along with other sequence alignment tools (e.g. SSEARCH and FASTA; Pearson, 2000), create local alignments between similar sequences using scoring matrices. Scoring matrices assign a similarity score to each pair of aligned amino acids based on the probability that the amino acid transition has occurred more often through evolution than by chance. Amino acid replacements that are common through evolution are assigned high similarity scores, while rare replacements are assigned negative scores. Scoring matrices have an implicit evolutionary model, which allows different matrices to target different evolutionary distances (Dayhoff *et al.*, 1978; Henikoff and Henikoff, 1992; Altschul, 1991; Muller *et al.*, 2002). Scoring matrices that target long evolutionary times (deep scoring matrices) allow more amino acid substitutions and gaps, while shallower matrices favor higher sequence identity and have higher gap penalties. The scoring matrix dictates the local alignment boundaries; increasing or decreasing the length of the optimal local alignment reduces the total alignment score. Likewise, changing the scoring matrix can result in a different alignment.

Ideally, a local alignment of homologous domains in different sequence contexts will align every residue in the homologous region, and no residues outside the domain boundaries, so that the alignment boundaries reflect the domain boundaries. Over-extended alignments include additional sequence from outside the homologous domain boundaries. For example, in Fig. 1 artificial, randomly shuffled, sequence from the query appears to be homologous to a real protein.

In this paper we show that scoring matrices have preferred alignment identities and alignment lengths, and that BLOSUM62 can produce over-extended alignments, most often between sequences with >33% identity. We also show that using the correct

*to whom correspondence should be addressed

scoring matrix can produce more accurate alignment boundaries. Finally, we show that we can produce more accurate alignment boundaries, even without true domain boundary knowledge, by using the initial BLOSUM62 alignment identity to specify a more appropriate scoring matrix.

2 METHODS

2.1 Construction of the RPD2 Dataset

Selecting families for RPD2. For this study, we built an updated version of the RefProtDom (RPD) protein database (Gonzalez and Pearson, 2010b) initially used to characterize alignment over-extension with PSI-BLAST, using protein domains and sequences annotated in Pfam version 26 (Punta *et al.*, 2012). From 13,672 initial Pfam version 26 families, 136 families were selected that met the following criteria: (i) model length (>200 residues); (ii) available structure; (iii) family size (>100 members); (iv) taxonomic diversity (presence in two of three kingdoms of life with the second most abundant kingdom having at least 15% as many the members as the most abundant). While most Pfam domain families can be represented by a single Hidden Markov Model (HMM), some very diverse families require multiple HMMs. When this occurs, the related domain families are grouped into Pfam clans. Protein domains belonging to the same Pfam family or Pfam clan are homologous to each other. Only a single family from any one clan was included, and then only if the family model lengths of the HMMs in the clan differed by less than two-fold. Of the 136 families selected, 56 were members of clans. Four RNA polymerase families were excluded because they have a complex and inconsistent domain organization.

Selecting sequences for the RPD2 library. For each of the RPD2 families, up to 5,000 non-viral, full-length (>80% of Pfam model length) domains were randomly selected. The unique protein sequences from which the domains came were then identified and included in the RPD2 library. Low complexity regions were lowercase masked by pseg and stored in FASTA format. Because many of these sequences contained domains other than the identifying domain, the final RPD2 library contains 1,837 families ranging in membership from 7,063 examples of the domain to 1. In total, the RPD2 library contains 499,058 domains from 282,742 different protein sequences.

Creating query sets for RPD2. For each RPD2 family, 10 non-viral, full-length examples of the domain were randomly selected. These domain sequences were used as queries against the RPD2 library. Searches were performed with SSEARCH version 36.3.6. The example of the domain that was able to find the largest number of the RPD2 library domains with an E()-score $\leq 10^{-3}$ was selected to be that family's query sequence. Each selected domain was embedded in the center of shuffled sequence with the same length and amino-acid composition as the original domain.

2.2 Database searches and scoring matrices

Searches were performed using BLASTP version 2.2.27+ (Camacho *et al.*, 2009) or SSEARCH version 36.3.6 (Pearson, 2000). A SSEARCH comparison of 136 query sequences against the 282,742 sequence RPD2 library took about 2 minutes on a 48 core machine. Bit scores, sequence identity, expectation values, and alignments were calculated by the search algorithm. All alignments had an E()-score $\leq 10^{-6}$ with a domain originally annotated by Pfam. Two types of scoring matrices were evaluated: the BLOSUM62 routinely used with BLASTP, and the VTML matrices described by Muller *et al.* (2002). For the VTML matrices, the gap penalties described by Reese and Pearson (2002) were adjusted to produce a smooth mean identity transition. The gap penalties used for each matrix are shown in Table 1.

2.3 Boundary accuracy

Boundaries for each alignment were known because the query domain was embedded in shuffled sequence. Alignments that extend outside of the embedded domains into the shuffled sequence are over-extended. Alignments that fail to extend to the domain boundaries are incomplete. Alignment boundaries within ± 10 residues of the embedded domain boundary are considered correct. The beginning and end of the alignments were evaluated independently, and the difference between the alignment boundaries and the embedded domain boundaries was calculated in number of residues. Incomplete alignments had negative boundary errors and over-extended alignments had positive boundary errors.

Table 1. Scoring matrices, gap penalties, and mean identity, entropy, and alignment length

Matrix	Open	Extend	Identity*	Entropy*	Length*
BLOSUM50	-10	-2	26%	0.24	178
BLOSUM62	-11	-1	30%	0.45	95
VT160	-12	-2	25%	0.28	155
VT140	-10	-1	31%	0.51	88
VT120	-11	-1	34%	0.63	67
VT100	-10	-1	40%	0.80	54
VT80	-11	-1	41%	0.82	54
VT40	-12	-1	65%	2.0	20
VT20	-15	-2	85%	3.3	11
VT10	-16	-2	93%	3.8	10

*Means measured from 136 random sequence searches (Fig. 3).

2.4 Sub-alignment scoring

SSEARCH from FASTA version 36.3.6 can provide location, identity, and score values for non-overlapping subsections of any alignment. In this study we annotated the embedded domain and non-domain regions in each query, which provided the score and identity for the homologous correct alignment, even if the alignment was over-extended. For over-extended alignments, the identity and score of the shuffled sequence that was included in the alignment was also calculated.

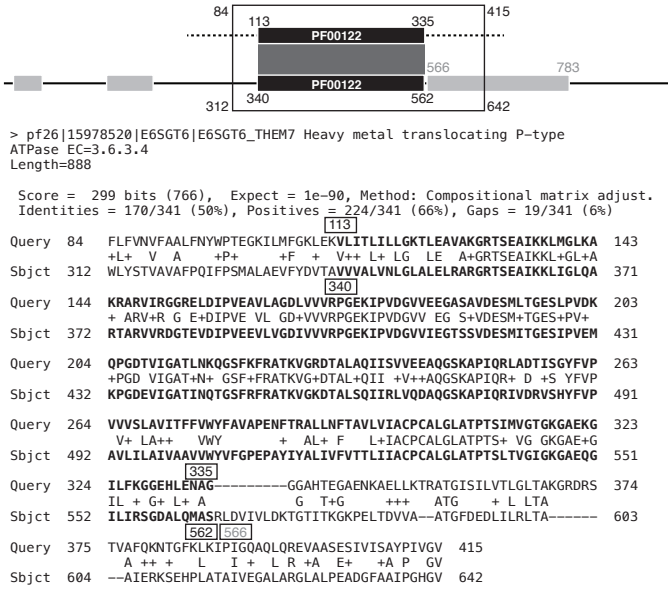
2.5 Scoring matrix adjustment

Alignments with greater than 36% identity were realigned using a series of VTML matrices. The new matrix was selected based on the BLOSUM62 identity given in Table 2.

3 RESULTS

3.1 Homologous over-extension

Deep scoring matrices can produce inaccurate alignment boundaries. Fig. 1 shows an example of an over-extended alignment created by BLASTP. The query was constructed using an E1-E2 ATPase (PF00122) domain from B0TE74_HELMI surrounded by shuffled sequence (dashed lines). This domain is homologous to the E1-E2 ATPase domain, also labeled PF00122, in the library sequence. The PF00122 domain extends from position 113 to 335 in the query. Any alignment that includes sequence from the query outside of the embedded domain includes shuffled sequence that is not homologous to the library sequence. In this example, the alignment extends from position 84 to 415 in the query, incorporating 109 residues of shuffled sequence or 33% of the total alignment length. The library sequence, like many proteins, consists of multiple domains. The alignment between these two sequences falsely indicates that shuffled sequence in the query is homologous to a neighboring Hydrolase (PF00702) domain in the library. BLASTP reports that the aligned sequences are 50% identical,




```
> pf26|15978520|E6SGT6|E6SGT6_THEM7 Heavy metal translocating P-type
ATPase EC=3.6.3.4
Length=888

 Score =  299 bits (766),  Expect = 1e-90, Method: Compositional matrix adjust.
 Identities = 170/341 (50%), Positives = 224/341 (66%), Gaps = 19/341 (6%)
                                                  113
Query  84   FLFVNVFAALFNYWPTEGKILMFGKLEKVLITLILLGKTLEAVAKGRTSEAIKKLMGLKA  143
            +L+  V  A    +P+     +F  +  V++ L+ LG  LE  A+GRTSEAIKKL+GL+A
Sbjct  312  WLYSTVAVAFPQIFPSMALAEVFYDVTAVVVALVNLGLALELRARGRTSEAIKKLIGLQA  371
                                              340
Query  144  KRARVIRGGRELDIPVEAVLAGDLVVVRPGEKIPVDGVVEEGASAVDESMLTGESLPVDK  203
            + ARV+R G E+DIPVE VL GD+VVVRPGEKIPVDGVV EG S+VDESM+TGES+PV+
Sbjct  372  RTARVVRDGTEVDIPVEEVLVGDIVVVRPGEKIPVDGVVIEGTSSVDESMITGESIPVEM  431

Query  204  QPGDTVIGATLNKQGSFKFRATKVGRDTALAQIISVVEEAQGSKAPIQRLADTISGYFVP  263
            +PGD VIGAT+N+ GSF+FRATKVG+DTAL+QII +V++AQGSKAPIQR+ D +S YFVP
Sbjct  432  KPGDEVIGATINQTGSFRFRATKVGKDTALSQIIRLVQDAQGSKAPIQRIVDRVSHYFVP  491

Query  264  VVVSLAVITFFVWYFAVAPENFTRALLNFTAVLVIACPCALGLATPTSIMVGTGKGAEKG  323
             V+ LA++   VWY       +  AL+ F   L+IACPCALGLATPTS+ VG GKGAE+G
Sbjct  492  AVLILAIVAAVVWYVFGPEPAYIYALIVFVTTLIIACPCALGLATPTSLTVGIGKGAEQG  551
                                  335
Query  324  ILFKGGEHLENAG----------GGAHTEGAENKAELLKTRATGISILVTLGLTAKGRDRS  374
            IL + G+ L+ A            G  T+G      +++   ATG    + L LTA
Sbjct  552  ILIRSGDALQMASRLDVIVLDKTGTITKGKPELTDVVA--ATGFDEDLILRLTA------  603
                                  562 566
Query  375  TVAFQKNTGFKLKIPIGQAQLQREVAASESIVISAYPIVGV  415
               A ++ +   L   I +  L R +A  E+   +A P  GV
Sbjct  604  --AIERKSEHPLATAIVEGALARGLALPEADGFAAIPGHGV  642
```

Fig. 1. Homologous overextension. BLASTP with BLOSUM62 was used to create an alignment between a RPD2 query and a homologous sequence from the RPD2 library. The raw BLASTP output and a schematic of the sequences are shown. Homologous domains in the query (top) and subject (bottom) sequence are represented by black boxes. Light grey boxes in the library sequence indicate other domains. The embedded domain in this query is from B0TE74_HELMI and sequence in the query outside of the embedded domain is shuffled. Black numbers show the homologous domain boundaries in both the schematic and raw BLASTP output (in boxes); grey numbers indicate the boundaries of the neighboring domain. The boundaries of the alignment are given by the open box while the correct alignment is represented by the dark grey box in the schematic.

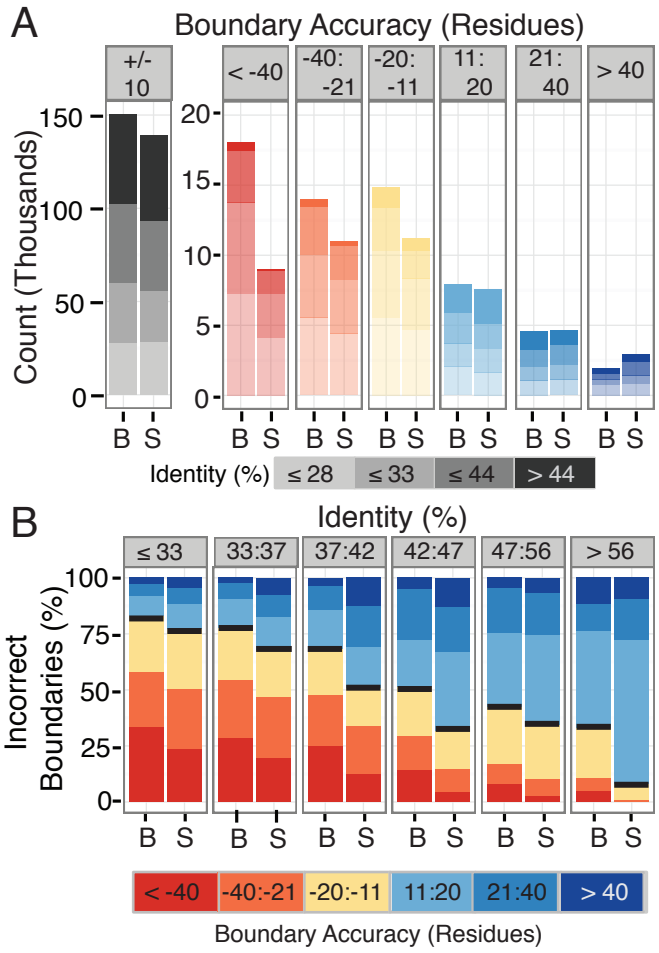


Fig. 2. Boundary accuracy and sequence identity. Using the RPD2 embedded domain queries and sequence library, pairwise protein sequence alignments were calculated with BLASTP (B) and SSEARCH (S) using BLOSUM62. Boundary accuracy was measured for both the beginning and end of alignments between known homologs with E()-score $\leq 10^{-6}$ as detailed in Methods. Alignment inaccuracy of <-10 residues indicates an incomplete alignment; >10 residues is considered over-extension. In panel (A), alignment identities were divided into quartiles. The data from the searches was binned by boundary accuracy (top) and sequence identity (color). In panel (B), incorrect alignment boundaries were isolated and alignments were divided into 6 identity bins. The boundary accuracy is given by the color of the bar. Identity bins are inclusive at the maximum.

but the homologous region is 64.1% identical while the non-homologous flanking regions are 23% identical. The homologous region contributes 83% of the bit score (248.2 bits) and the non-homologous region only contributes 17%. This imbalance in the contributions of homologous compared to non-homologous regions to both alignment identity and score is a hallmark of over-extended alignments.

3.2 Over-extension occurs more frequently in alignments with higher sequence identity

To understand how often incorrect alignment boundaries occur, searches were performed with both BLASTP and SSEARCH, using BLOSUM62 (BL62) with the RPD2 query set and library. Each alignment boundary was measured and the results were divided into seven bins ranging from extremely incomplete (<-40 residues, i.e., more than 40 residues missing) to extremely over-extended (>40 residues added; Fig. 2A). While most of the alignment boundaries were within 10 residues of the embedded domain boundaries (71% BLASTP, 75% SSEARCH), BLASTP and SSEARCH also created incorrect alignment boundaries. Of the boundaries measured, 22% of BLASTP boundaries were incomplete and 7% were over-extended, aligning random sequence with real protein residues. Seventeen percent of the SSEARCH boundaries were incomplete and 8% were over-extended. Alignment identity was divided into quartiles. Each identity quartile shows similar representation within the group of "correct" alignment boundaries (within ±10 residues of the embedded domain). In contrast, incomplete alignment boundaries are more common in low identity alignments while over-extended alignment boundaries are more common in high identity alignments. Most incomplete alignment boundaries (73% for BLASTP, 76% for SSEARCH) were from alignments in the lowest two identity quartiles. The opposite is true for over-extended alignments, where most had identities in the top two quartiles (52% for BLASTP, 54% for SSEARCH). When incorrect alignments are examined independently, the percentage of the boundaries that are over-extended increases with identity (Fig. 2B).

Fig. 2 reports incomplete and over-extended alignment boundaries for the 397,123 homologs that were identified by BLASTP and SSEARCH. Because RPD2 was built from diverse domain families, most of these homologs are very distant, with a median identity of 33%. In practice, one rarely examines every significant match, so we also counted incomplete and over-extended boundaries for the top 100 significant hits with each query. For the top 100 hits, the median alignment identity increases to 52%. In this more closely related set, the percentage of over-extended alignments increases to 8% for BLASTP and 10% for SSEARCH and incomplete alignment decreases to 8% and 5% respectively.

Incomplete alignments can occur when homologous domains are evolutionarily distant, so that the alignment captures only the most conserved regions of the homology. This contrasts with traditional false-negatives, where the homology is missed altogether. In the traditional case, the reduced sensitivity of pairwise sequence comparisons compared with model-based (PSI-BLAST, PSI-SEARCH, HMMER) or structure based methods is well recognized (Pearson and Sierk, 2005). Incomplete alignments are another example of inadequate alignment sensitivity.

Over-extension, while recognized in pairwise genomic alignments (Chao *et al.*, 1993), had not been systematically measured in pairwise protein alignments. Missed homologs can be identified using transitive homology, protein family models, or structures. But strategies for removing non-homologous sequence from pairwise protein alignments have not been described.

3.3 Scoring matrices, identity and alignment length

Alignment over-extension often results from a mismatch between the evolutionary distance between the homologous sequences and the target identity of the scoring matrix used in the alignment. Unlike global sequence alignments, which use the full length of each sequence, the scoring matrix determines local alignment boundaries. To understand how different scoring matrices produce different alignment boundaries, we used shuffled sequences as queries against the RPD2 library.

"Deeper" scoring matrices (scoring matrices targeted to more evolutionary change) produce longer, less identical alignments by chance, while "shallower" scoring matrices produce shorter, higher identity alignments (Fig. 3). Here, the same 136 shuffled queries were used with each matrix, so the resulting trends in identity and alignment length reflect the average properties of the matrices themselves. The target identities with gaps are lower, and the alignment lengths longer, than the values estimated from the scoring matrix alone. Remarkably, the entropies calculated analytically from the scoring matrix alone track closely between the gapped and un-gapped empirical mean entropies. Including gaps (black boxes) makes scoring matrices "deeper," thus lowering identity and increasing alignment length compared to the same matrix without gaps (grey circles). Different scoring matrices can produce different alignment boundaries.

3.4 Selecting the correct scoring matrix gives correct domain boundaries

To illustrate how "correct" scoring matrices—scoring matrices with target identities that match the evolutionary distance of the homologous domains—improve accuracy, we examined alignment boundary changes with different scoring matrices. Beginning

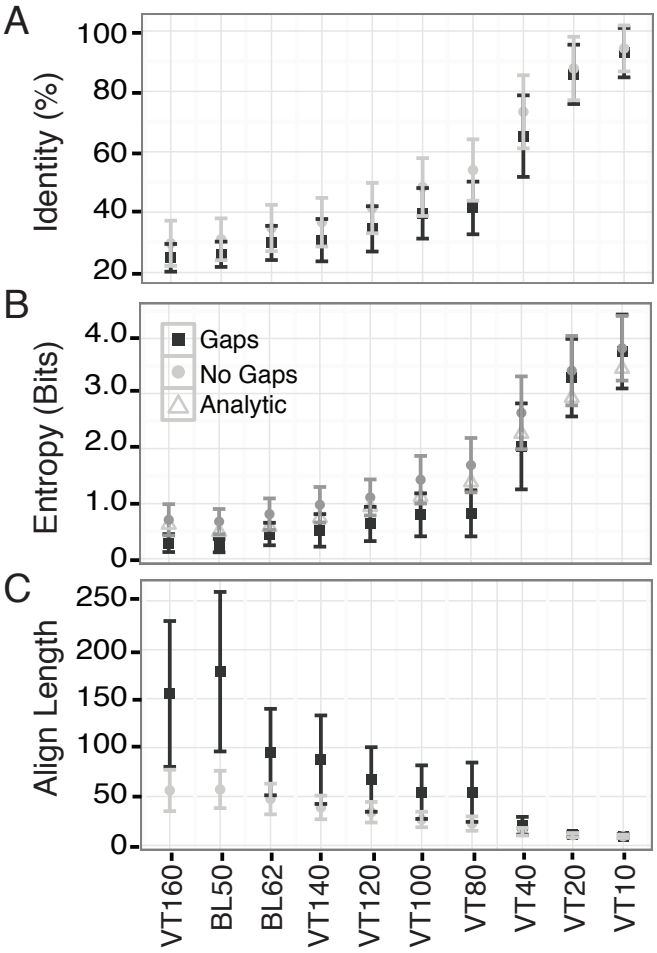


Fig. 3. Scoring matrix target identity, entropy, and alignment length. Queries were constructed from 136 shuffled protein domains. SSEARCH was used to search against the RPD2 library with these shuffled queries using either the gap penalties given in Table 1 (black squares) or gap penalties of -1000/-1000 for open/extend (grey circles), which effectively creates alignments with no gaps. The identity and alignment length from the highest scoring alignment was selected from each query. The (A) mean identity, (B) mean entropy, and (C) mean alignment length, is given by the point and the standard deviation is indicated by the error bars for each scoring matrix. The analytical entropy calculated from the scoring matrix is shown as open triangles in panel (B).

with 16,640 over-extended alignments, we tracked the boundary accuracy produced by six VTML matrices (VT) with increasing target identity (Fig. 4). The alignment with the smallest cumulative difference between the embedded domain boundaries and the alignment boundaries was identified, and ten re-alignments from each of the VT scoring matrices were randomly selected. The maximum boundary errors for both the initial BLOSUM62 and final best alignment are shown in Fig. 4. All of the re-alignments corrected the over-extended boundary to within ± 10 residues of the embedded domain, producing alignments with higher identities. As the identity of the initial alignment increases, the target identity of the matrix that produces the corrected alignment also increases. However, the matrix required did not correlate with the amount of

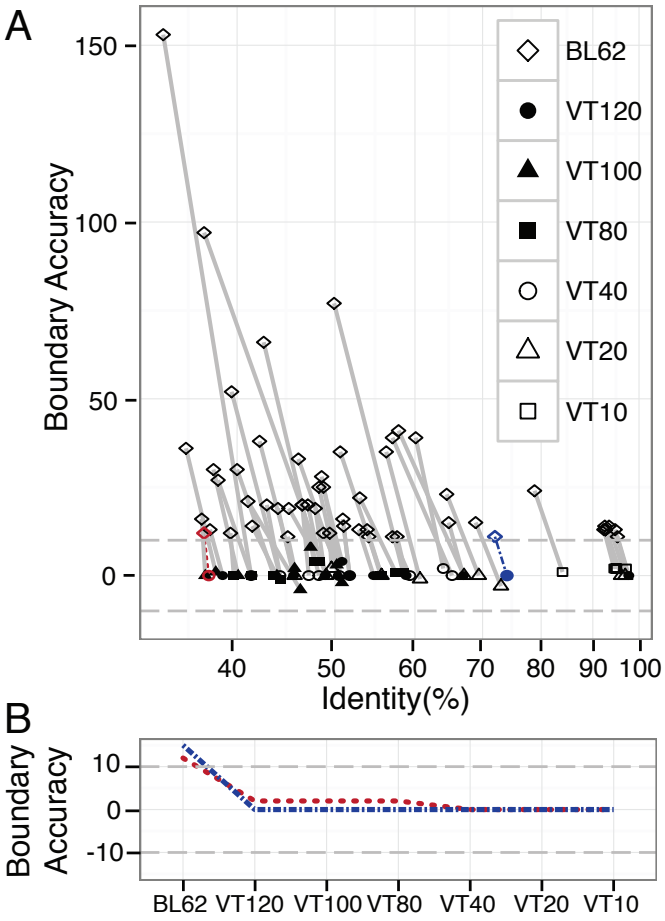


Fig. 4. Selecting the scoring matrix that creates the best alignment. (A) Sequence pairs with >33% identity and over-extended alignment boundaries were selected from the results of the similarity search performed using SSEARCH with BLOSUM62. Each sequence pair was realigned using VT120,100,80,40,20 and VT10 (Table 1). Boundary accuracy was calculated for each alignment and the alignment with the smallest cumulative difference between the embedded domain boundaries and the alignment boundaries was selected. Symbol shape and color (black, open) indicate the scoring matrix used for the alignment; lines connect alignments between the same sequence pairs. (B) Maximum boundary inaccuracy across every scoring matrix for two sequence pairs in (A) is shown. The rounded dashed line to the left in panel (A) and the higher line between VT120 and VT80 in panel (B) shows a low identity alignment corrected by VT40; the square dash-dot line to the right in panel (A) and flat between VT120 and VT10 in panel (B) shows a high-identity alignment corrected by VT120.

over-extension in the original BLOSUM62 alignment in this data set. Nor was there any correlation in alignments that used alternate shuffling strategies for the embedded domains.

In general, lower target identity matrices (VT120, VT100, VT80) correct lower identity alignments (the filled symbols tend to be on the left of the final distribution) and higher target identity matrices (VT40, VT20, VT10) correct higher identity alignments (the open symbols tend to be on the right). But this is not always the case; sometimes a high identity alignment is corrected by a distant matrix (dash-dot line), and vice versa (rounded-dash line).

Anomalous matrices can correct over-extension because alignment boundary correction is robust to matrix selection. Fig. 4B shows two extreme examples, a deep matrix (VT120) correcting a high-identity alignment (dash-dot line), and a shallow matrix (VT40) correcting a low-identity alignment (rounded-dash line). In both cases, a wide range of scoring matrices correct the alignment, including a matrix at the predicted target identity (for the red low-identity alignment, VT120, VT100, and VT80 produce an alignment that is off by two residues, while VT40 is perfect). The robustness of boundary correction to scoring matrix choice allows us to approximate the "correct" alignment identity from the initial (possibly over-extended) BLOSUM62 identity.

Since high identity alignments tend to be corrected by shallow scoring matrices while lower identity alignments can be corrected by less shallow scoring matrices (Fig. 4), we attempted to correct BLOSUM62 alignments using the scoring matrices and thresholds shown in Table 2.

Table 2. Identity required to re-align using each scoring matrix.

Matrix	Identity Range
VT120	36-50%
VT100	50-60%
VT80	60-70%
VT40	70-80%
VT20	80-85%
VT10	>85%

Values are inclusive at the maximum for each matrix.

Forty-seven percent of over-extended boundaries came from alignments with >36% identity and therefore were candidates for the re-alignment algorithm. Of the over-extended boundaries that could be re-aligned, 97% had reduced over-extension with 86% of the over-extended boundaries moving within ±10 residues of the embedded domain boundaries. Overall, including over-extended alignments that were not re-aligned, the total amount of over-extension was reduced from 8% to 5%.

While the scoring matrix identity thresholds in Table 2 dramatically decrease over-extension errors they can also produce incomplete alignments (Fig. 5). In contrast to Fig. 4, where we selected the most accurate alignment, Fig. 5 shows the results of realignment based solely on the identity of the initial BLOSUM62 alignment (the thresholds in Table 2). Looking at all alignments with >36% identity, 16,411 alignment boundaries changed accuracy bins. Of the alignment boundaries that changed accuracy bins, 68% moved from being over-extended (>10 residues, blue colors) to within ±10 residues, while 20% moved from being within ±10 residues or over-extended to incomplete. Most (73%) of the re-aligned incomplete alignment boundaries fall into the $-20 : -11$ bin (orange). The most over-extended alignments (>40 residues, Fig. 2) decreased by 2,217 alignment boundaries, while the most incomplete alignments increased by 399 boundaries. The final distribution of all alignment boundaries had 7,863 more boundaries within 10 residues of the embedded domain boundary and 3,189 additional incomplete boundaries, or 2.5 additional boundaries within ±10 residues for each additional incomplete boundary.

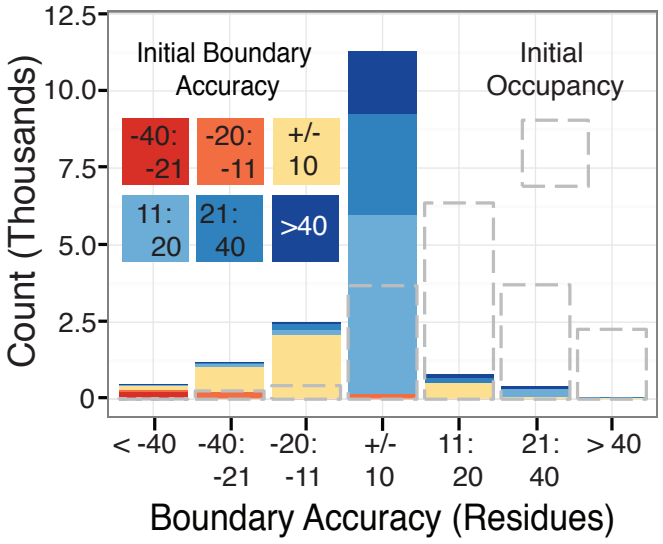


Fig. 5. Re-Alignment Algorithm Results Only results from sequence pairs that were re-aligned by the algorithm are shown. Grey dashed bars indicate the initial boundary accuracy before re-alignment; colored bars indicate the final distribution of alignment boundary errors. The colors in the final distribution bars show the original accuracy before re-alignment.

Alignment boundary correction is much more effective when applied to the highest scoring alignments. Focusing on the top 100 alignments from each query, 83% of the over-extended boundaries were from alignments with >36% identity of which 90% moved within 10 residues of the embedded domain boundaries reducing the amount of over-extension from 10% to 3%. The top 100 alignments produced many fewer incomplete alignments; 1,645 boundaries moved to within 10 residues of the embedded domain while only 233 boundaries became worse than <-10 residues incomplete, a ratio of 7 corrected boundaries for each additional incomplete boundary.

4 DISCUSSION

Mismatches between the sequence identity of aligned homologous domains and the target identity of the scoring matrix used to produce the local sequence alignment can lead to over-extended alignments (Fig. 1, Fig. 2). Similarity scoring matrices have preferred alignment lengths and identity. Deep scoring matrices create longer alignments and have lower target identity compared to shallower matrices (Fig. 3). Alignments created by BLOSUM62, most often between sequences with higher identity (>33%), can extend past the boundaries of the homologous domain to include non-homologous sequence (Fig. 2). Using a shallower scoring matrix that targets the correct sequence identity can correct over-extension (Fig. 4). Predicting the scoring matrix that will lead to a better alignment, using initial (possibly over-extended) identity given by BLOSUM62, can correct over-extended alignments. In our RPD2 database, 37% of over-extended alignments were corrected to within ±10 residues, or 86% of the alignments with high enough identity (>36%) to be considered for re-alignment. However, re-alignment has a cost; a fraction of correctly aligned domains are incompletely re-aligned.

The observation that "deep" scoring matrices produce over-extended alignments between domains that are less evolutionarily distant (have higher identity) than the scoring matrix target identity is not surprising, though the relationship between alignment boundaries (in contrast to internal alignment accuracy) and scoring matrices has not been extensively studied. Traditional internal alignment accuracy decreases as evolutionary distance increases; very different sequences are difficult to align accurately. In contrast, alignment over-extension occurs most often when closely related sequences are aligned, and thus becomes more frequent as sequence databases grow.

As log-odds matrices, every scoring matrix has a target evolutionary distance, or percent identity, which can be approximated from the homologous replacement frequencies that are the numerator of the log-odds ratio (Altschul, 1991). As evolutionary distance and the number of replacements increase, the replacement frequencies for identities decrease and the non-identical replacement frequencies increase, which reduces the target identity of the matrix when aligning random sequences (Fig. 3). Over-extension occurs when a scoring matrix that models a longer period of evolution (a deeper scoring matrix) by allowing more mutations is used to align sequences with less evolutionary change. A deep scoring matrix produces a less identical alignment because it accepts more amino-acid replacements. Gap penalties also modify alignment length and identity; increased gap penalties produce shorter, higher identity alignments while lower gap penalties produce longer, lower identity alignments (Fig. 3). Lower mismatch penalties and lower gap penalties in deep matrices allows the local alignment algorithm to add additional identities that are occurring by chance from non-homologous sequence for the sake of modest increases in score. Thus, in Fig. 1, 83% of the score was produced by 67% of the alignment. Over-extended alignments are locally optimal, but they are not biologically correct.

RPD2 was designed to simulate the most common similarity search; searches against full-length proteins in a comprehensive sequence database. RPD2 sequences were selected from the set of sequences annotated by Pfam release 26, which samples both SwissProt and Trembl protein sequences. The RPD2 library is large (528,742 sequences) and diverse. Queries were engineered from long (>200 residues) protein domains, allowing BLOSUM62 searches to identify distant homologs. These domains are surrounded by shuffled protein sequence, providing known alignment boundaries. Alignments that extend into the flanking random sequence are thus guaranteed to be non-homologous.

Our initial searches with 136 independent embedded domain queries produced both incomplete alignments (22% BLASTP, 17% SSEARCH, both with BLOSUM62) and over-extended alignments (7% BLASTP, 8% SSEARCH). Incomplete alignments reflect the reduced sensitivity of pairwise alignment compared with the Hidden Markov-Model base methods used to annotate the Pfam domains in RPD2, and the fact that in the diverse set of homologous RPD2 domains, half of the detectable homologs share less than 33% sequence identity.

In characterizing more than 2 × 200,000 alignment boundaries in the 136 query domain searches, we consider far more distant alignments than would typically be examined during the genome

annotation process, where sequences sharing at least 40% identity might be used to transfer annotation. Restricting the analysis to the top 100 significant hits for each query increases the median alignment identity to 53%, which in turn decreases incomplete alignments to 5%, and increases the over-extension to 10%. Restricting the analysis to the top 25 homologs further decreases incomplete alignment to 2%, while increasing over-extension to 11%. When the thresholds in Table 2 are used to correct the top 100 alignments for each query, over-extension is corrected 73% of the time, while incomplete alignments are produced only about 10% of the time. For the top 25, over-extension is corrected 86% of the time, while alignments become incomplete only 1% of the time.

We believe our estimates of alignment over-extension (7 – 10% of alignments) are conservative, both because sequence databases are growing, allowing similarity searches to identify closer homologs, and because many proteins are comprised of multiple domains. In this study, we examine alignment over-extension from a single domain. Many proteins contain multiple domains separated by non-homologous regions; proteins that contain multiple widely dispersed common domains, like Ankyrin, fn3, or SH3 domains, will have many more chances to over-extend across non-homologous regions.

Although we understand why high identity alignments might over-extend when aligned with low target-identity scoring matrices like BLOSUM62, matrix/target-identity mismatch only accounts for about half of the over-extensions we observed. In our diverse sequence set, 53% of over-extensions occur in alignments that are $<$36% identical. Unfortunately, we cannot predict which lower-identity alignments will over extend. The amount of over-extension does not correlate well with the difference between alignment identity and scoring matrix target identity. Likewise, over-extension does not occur significantly more often in domains that have more identity at their ends. The increased frequency of over-extended boundaries in alignments between high identity sequences (Fig. 2B) is the only meaningful trend that we identified.

In contrast to high-identity over-extension, where the difference in target-identity between the homologous region and the scoring matrix can explain over-extension, low-identity over-extension may simply reflect the propensity of deep matrices to produce long alignments, even between unrelated sequences (Fig. 3). The long alignments in Fig. 3 are not statistically significant, but when they occur by chance near a (low-identity) homologous domain, they can contribute to over-extension. Over-extension occurs more frequently in higher identity alignments because of target-identity mismatch, but the majority of over-extension we measured occurs by chance in low-identity alignments, because most of our alignments are low-identity.

In this study, we have focused on over-extension in pairwise alignments, because pairwise similarity searches are widely used to annotate newly sequenced genomes. Alignment over-extension also occurs with model-based searches like PSI-BLAST; indeed, we initially identified over-extension as the major cause of model contamination with PSI-BLAST (Gonzalez and Pearson, 2010a). Our strategy for reducing over-extension — re-alignment with a more correct scoring matrix — is most easily applied to pairwise alignment, because a traditional non-position specific scoring matrix like BLOSUM62 or VT120 has an easily characterized target identity and the alignment between two sequences has a natural evolutionary distance. It is more difficult to interpret the "distance" between a sequence and a position-specific scoring matrix or HMM, and it is unclear how such models might be scaled to reduce over-extension.

The expansion of modern protein databases has led to an increase in the identification of higher identity homologs. Accurate function prediction requires a higher level of sequence identity and an accurate alignment, two factors that are at odds with deep scoring matrices. With modern comprehensive databases, it is common to identify many homologs that are $>$40% identical. In our diverse RPD2 protein set, the median sequence identity for the top 100 homologs was 53%, much higher than the target identity range for BLOSUM62. With more high identity homologs and increased sequence and structural annotation, pairwise alignments can provide essential insights to the function of novel proteins, but only if the alignment boundaries are accurate.

ACKNOWLEDGEMENTS

The authors thank R. Clark and F. Elliot for technical support, and A. J. Mackey for useful discussions. We thank a reviewer for suggesting an explanation for low-identity over-extension.

REFERENCES

Altschul, S. F. (1991). Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **219**, 555–65.

Altschul, S. F., Madden, T. L., Schäffer, 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 Acid Res.*, **25**, 3389–3402.

Arslan, A. N., Eğecioğlu, Ö., and Pevzner, P. A. (2001). A new approach to sequence comparison: normalized sequence alignment. *Bioinformatics*, **17**, 327–337.

Brenner, S. E., Chothia, C., and Hubbard, T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA*, **95**, 6073–6078.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). Blast+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Chao, K.-M., Hardison, R. C., and Miller, W. (1993). Locating well-conserved regions within a pairwise alignment. *Comput. Applic. Biosci.*, **9**, 387–396.

Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, volume 5, pages 345–358. National Biomedical Research Foundation, Washington DC.

Gonzalez, M. W. and Pearson, W. R. (2010a). Homologous over-extension: a challenge for iterative similarity searches. *Nucleic Acids Res.*, **38**, 2177–2189.

Gonzalez, M. W. and Pearson, W. R. (2010b). Refprotdom: A protein database with improved domain boundaries and homology relationships. *Bioinformatics*, **26**, 2361–2361.

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

Muller, T., Spang, R., and Vingron, M. (2002). Estimating amino acid substitution models: A comparison of dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Mol. Biol. Evol.*, **19**(1), 8–13.

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

Pearson, W. R. (2000). Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.

Pearson, W. R. and Sierk, M. L. (2005). The limits of protein sequence comparison? *Curr Opin Struct Biol*, **15**, 254–260.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., Bateman, A., and Finn, R. D. (2012). The pfam protein families database. *Nucleic Acids Res*, **40**, D290–D301.

Reese, J. T. and Pearson, W. R. (2002). Empirical determination of effective gap penalties for sequence comparison. *Bioinformatics*, **18**, 1500–1507.