

MAXIMAL INFORMATION FROM LOCAL ALIGNMENTS

Lauren Jennifer Mills
Austin, Texas

B.S. Biology Centenary College of Louisiana, 2007

A Dissertation presented to the Graduate Faculty of the University of Virginia
in Candidacy for the Degree of Doctor of Philosophy

Molecular, Cell and Developmental Biology

University of Virginia
December, 2013

Abstract

Accurate identification of homologs, through sequence similarity searching programs like BLAST, is central to converting genome sequence data to biological knowledge. BLAST, FASTA, and other widely used search programs use local alignments to identify homologous sequences based on shared domains, but the boundaries of local alignments reflect both the signal from homology and the intrinsic properties of the alignment scoring matrix. Reliable identification of homologous domains requires sensitive alignment methods, accurate statistical estimates, and accurate alignment boundaries. Matrices that produce sensitive searches can also produce inaccurate alignments for more closely related homologs. Past improvements in search strategies focussed on search sensitivity and statistical accuracy, but largely ignored boundary accuracy. Homologous overextension, a boundary error that occurs when two homologous domains are aligned, but the alignment extends beyond the ends of the domains, can propagate inaccurate functional predictions and contaminate models used in more sensitive similarity searches.

In this thesis, I discuss the theoretical and empirical basis for homologous overextension. In Chapter 1, I outline the properties of local similarity scoring matrices that can produce alignment overextension. In Chapter 2, I show that overextension occurs in 8% of alignments in comprehensive searches, increasing to 10% for the 100 most similar alignments. About half of this overextension occurs because of a mismatch between the alignment identity of the homologous domain and the target identity of the scoring matrix used in the initial alignment and more than 85% of this high-identity alignment overextension can be corrected by shifting to the appropriate scoring matrix. In Chapter 3, I consider alignment over extension in other contexts and summarize additional strategies for identifying over extension. Alignment accuracy is central to effectively exploiting our growing knowledge about structure-function relationships, active sites, and variant phenotypes. Future characterizations of alignment methods should examine both internal and alignment boundary accuracy.

Contents

1		1
1.1	Motivation	1
1.2	Inferring Homology	2
1.3	Sequence Alignment Boundaries	5
1.4	The Algebra of Similarity Scoring Matrices	7
1.5	Calculating Substitution Frequencies	8
1.6	Scoring Matrix Depth and Alignment Length	10
1.7	Homologous Over Extension	13
2		17
2.1	Abstract	17
2.1.1	Motivation:	17
2.1.2	Results:	17
2.2	Introduction	18
2.3	Methods	19
2.3.1	Construction of the RPD2 Dataset	19
2.3.2	Database searches and scoring matrices	21
2.3.3	Boundary accuracy	21
2.3.4	Sub-alignment scoring	21
2.3.5	Scoring matrix adjustment	22
2.4	Results	22
2.4.1	Homologous over-extension	22
2.4.2	Over-extension occurs more frequently in alignments with higher sequence identity	23
2.4.3	Scoring matrices, identity and alignment length	28
2.4.4	Selecting the correct scoring matrix gives correct domain boundaries	28
2.5	Discussion	32
3		38
3.1	Summary of Major Conclusions	38
3.2	Other Algorithms to Identify Non-Homologous Residues	39
3.3	Sub-Alignment Scoring and Incorporation of Outside Annotations	41
3.4	Complex Protein Architectures	42
3.5	Boundary Accuracy in DNA Alignments	42
3.6	Improving Iterative Search Strategies	45

3.7	Final Thoughts	46
-----	--------------------------	----

Chapter 1

1.1 Motivation

Sequence alignment is fundamental to understanding genome function, providing insights into biological function, disease processes, and the relationships between genomes. Sequence alignment also enables the examination of the evolutionary context that connects structure to function. The only way to trace a sequence's evolutionary history is by identifying sequences that share a common ancestor, or in other words, by identifying homologs.

Given the explosion of genome sequence data, sequence alignments are performed as similarity searches, which seek to identify homologous proteins and protein domains. Similarity searching combines two calculations: 1) a sequence alignment based on a scoring matrix and 2) a statistical estimate of how often the alignment could be expected by chance. Similarity search algorithms create alignments maximizing the similarity score between two sequences (step 1 above). However, similarity alone does not establish homology. Instead, we rely on accurate alignment similarity score statistics (E()-scores) to infer homology (step 2). The alignment, its similarity score, and the statistical significance can not be separated. The alignment sets up the comparison, and inferences taken from the statistics only apply to the regions of the sequences that are included in the alignment.

In this dissertation I will show that while similarity searches accurately identify homologs, the alignments produced during the search do not always reflect the true boundaries of the homologous region. I will then link the inclusion of non-homologous residues

to a mismatch between the evolutionary distance between the sequences and the evolutionary distance that the scoring matrix assumes. Finally, I will show how selecting a scoring system that better reflects the evolutionary distance between the sequences can correct inaccurate alignment boundaries leading to more accurate identification of homologs.

1.2 Inferring Homology

Understanding the evolutionary context of a biological sequence is a powerful first step towards understanding the sequence's function. Inferring homology is the only way to understand the evolutionary context of a specific sequence. Homologous sequences share a common origin (Koonin and Galperin, 2003; Wagner, 1989), i.e., they are all copies of one original sequence and share the same structure. The structure and function of a protein are strongly linked. The inference of shared structure is what makes identifying homologs a powerful tool for understanding novel proteins. When we infer that a human sequence is homologous to an *E. coli* sequence, these two sequences are both copies of an original sequence that existed more than two billion years ago in the common ancestor shared by humans and *E. coli*. These sequences have changed so slowly from the original sequence that they can be recognized as copies of one another even in their modern forms and still share the same structure.

Homology can be inferred by identifying sequences that share excess similarity, i.e., more similarity than would be expected by chance. The most parsimonious explanation for why two sequences share excess similarity is that they descended from (are copies of) a common sequence; that is, they are homologous. This raises the question, how much similarity is excess? Furthermore, what is the smallest amount of similarity that two sequences can share and still be considered to have more than would be expected by chance? Accurate statistics are needed to establish when a similarity score is large enough to be considered excess. A similarity score can be used to recognize excess similarity because the amount

of similarity seen between non-homologous sequences is indistinguishable from the similarity seen between randomly generated sequences. A similarity score that is unlikely to occur between randomly generated sequences (i.e., that is unlikely to occur by chance) is evidence for excess similarity and therefore homology (Karlin and Altschul, 1990; Pearson, 1996).

Using the expected distribution of similarity scores for a given search, statistical estimates can indicate if an alignment similarity score is expected to occur by chance. Scores from local alignments are known to approach an extreme value distribution (Karlin and Altschul, 1993) that can be calculated using the parameters of the search e.g., database size, length of the query sequence and scoring matrix used. Using the extreme value distribution for a given search, expectation scores (E()-scores) calculated for each similarity score indicate the likelihood that the similarity score (S) could occur by chance where $E = Kmn e^{-\lambda S}$, where λ adjusts for the scale of the scoring matrix, K adjusts the alignment length effect and m, n are the length of the query and length of the database, respectively (Altschul and Gish, 1996; Altschul *et al.*, 1994). The FASTA programs use an equivalent formulation that explicitly factors in the number of sequences in the database. Given a particular alignment and the corresponding similarity score, an E()-score of 10^{-4} indicates that this particular similarity score could be expected to happen by chance once every 10,000 searches, while a similarity score with an E()-score of 1 would happen (by chance) in every search. Similarity scores that have statistically significant E()-scores (less than 10^{-3} for a single search) are strong evidence of excess similarity and thus homology (Brenner *et al.*, 1998).

While E()-scores can be used as positive indicators of homology, a non-significant E()-score is not an indication of non-homology. Instead, it indicates a lack of evidence from which to infer homology. The ultimate evidence for both homology and non-homology is structural similarity. Structures diverge more slowly than sequences and are a more sensitive measure from which to infer homology. If proteins do not share the same structure

they are not homologous, because the sequences could have just as easily arisen independently. There are thousands of examples of sequences that do not share statistically significant similarity but do have the same structure and are therefore homologous.

A drawback of pairwise searching is that the search sensitivity is limited to the query's unique perspective of the evolutionary tree connecting the homologs. If the sequence of interest happens to be in a highly populated part of the tree, then pairwise searching will be able to infer many homologs. For protein families that exhibit large degrees of divergence, each sequence will be more isolated and pairwise methods will be unable to infer many homologs.

More sensitive methods than pairwise searching (i.e., one sequence against a database, as alluded to above) exist to infer homology via sequence alignment, though none of them are as sensitive as structural alignment. Alternative methods that use the transitive property of homology can link the perspectives of individual sequences together through shared inferred homologs. Simply put, the transitive property of homology states that if sequence A is homologous to sequence B, and the same region of sequence B is homologous to sequence C, then sequence A and C are also homologous in that shared region even if they do not share a statistically significant level of similarity. Additionally, methods that create position specific scoring matrices (Altschul *et al.*, 1997; Li *et al.*, 2012) and hidden Markov models (Eddy, 2011; Edgar, 2004; Koestler *et al.*, 2012) try to get beyond a single sequence's perspective of the tree by modeling the substitutions seen within the whole family of homologous sequences. In essence, both of these methods try to model the evolution of a single protein family by creating a similarity scoring system that is specific to a family instead of scoring only based on a single query. As with the use of the transitive property of homology, these methods allow for identification of homologs beyond the perspective of a single query.

1.3 Sequence Alignment Boundaries

The first homologs identified were from single domain protein families such as globins and cytochrome c (Fitch, 1976; McLachlan, 1971; Haber and Jr., 1970; Feng *et al.*, 1985). These single domain protein families could be aligned using the Needleman and Wunsch algorithm, which creates optimized global alignments (Needleman and Wunsch, 1970). Global alignment algorithms both offset proteins and insert gaps in the alignment to account for insertions and deletions to maximize the similarity seen between two full-length sequences. Global alignment boundaries are easily determined, as the alignment must extend to the ends of the sequences. Global alignments are most appropriate for single domain sequences, such as the alignment between Human and *Xenopus* hemoglobin seen in Figure 1.1A, or sequences that are closely related (e.g., most human and mouse proteins share global similarity).

In 1983, Doolittle discovered multi-domain proteins; proteins that shared domain homology, but were not homologous from beginning to end (Doolittle *et al.*, 1983). The local alignment algorithm created by Smith and Waterman was able to align individual domains by allowing the alignment to begin and end anywhere within the sequences based on where the most similarity existed (Smith and Waterman, 1981). Local alignment allows individual domains to be aligned regardless of their context in different protein sequences. For example, an alignment between Human cortactin (SRC8_HUMAN) and the *Xenopus* protein dbnl-b (DBNLB_XENLA) achieves statistical significance because the alignment can be limited to the local similarity that exists between the SH3 domains (Figure 1.1B) even though a vast majority (more than 85%) of the sequences are not homologous. The ability to create a local alignment allows these sequences to be accurately inferred as homologs ($E\text{-score} < 2.4 \times 10^{-27}$) though the single (SH3) domain they share. The boundaries of the alignment also limit the inference of homology to the single domain that these sequence share by excluding any other residues in the sequences from the alignment. That is, homology is only established within the boundaries of the alignment, and thus, domain

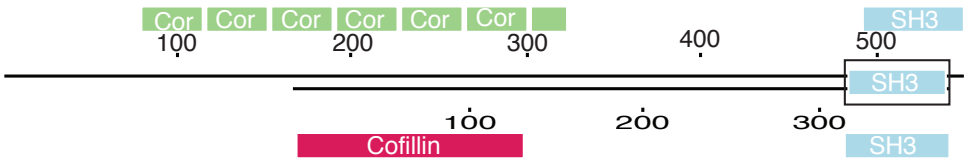
A) Global Alignment

```
>>sp|P02012|HBA1_XENLA Hemoglobin subunit alpha-1; Alpha-1-globin; (142 aa)
n-w opt: 414 Z-score: 314.4 bits: 64.1 E(13143): 3.3e-150
global/global (N-W) score: 414; 56.3% identity (81.0% similar) in 142 aa overlap (1-142:1-142)

      1      10      20      30      40      50      60      70      80
HUMAN MVLSPADKTNVKAAGKVGAGHAGEYGAELERMFLSFPTTKTYFPHFDSLHGSAQVKGHGKKVADALTNAVAHVDDMPNA
      :::  ::  ::::  ::::  :::::  :::  :  :::::  ::  :::  ::::::::::::::  ::::  ..
XENLA MLLSADDKKHIKAIMPAIAAHGDKFGGEALYRMFIVNPKTKTYFSPDFHHNSKQISAHGKKVVDALNEASNHLNDNIAGS
      1      10      20      30      40      50      60      70      80

      90      100      ]10      120      130      140
HUMAN LSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTISKYR
      .:  :::::  ::::  ::  :::  ::::  ::  .:  :::  :::::::::::::::
XENLA MSKLSDLHAYDLRVDPGNFPLLAHNILVVMNFPKQFDPATHKALDKFLATVSTVLTISKYR
      90      100      ]10      120      130      140
```

B) Local Alignment



```
>>sp|Q6GM14.1|DBNLB_XENLA Drebrin-like protein B (376 aa)
initn: 296 initl: 219 opt: 277 Z-score: 625.0 bits: 124.9 E(455806): 2.4e-27
Smith-Waterman score: 280; 66.7% identity (72.5% similar) in 69 aa overlap (481-548:315-374)

      490      500      510      520      530      540      550
sp|Q1  AEDSTYDEYENDLGITAVALYDYQAAGDDEISFDPDIIITNIEMIDGWWRGVC-KGRYGLFPANYVELRQ
      :::  .:  :  :::::  :  :::::  :  :::::  :  :  :::::  :  :::::  :
sp|Q6G AEDS-----GMCARALYDYQAADDTEISFDPDVIIQIEMIDGWWRGVAPSGHFGMFANYVELLE
      320      330      340      350      360      370
```

Figure 1.1: Global and Local Alignments. A) Global alignment using Needleman and Wunsch algorithm between HBA_HUMAN (142 residues) and HBA1_XENLA (142 residues). Colons indicate an identity and a period indicates a similarity with a score of 0 or greater. The alignment between these homologs is 142 residues long, 56.3% of the residues are identical and has a statistically significant E()score of 3.3e-150. B) Local alignment using Smith and Waterman algorithm and VTML40 scoring matrix between SRC8_HUMAN (550 residues) and DBNLB_XENLA (376 residues) shown as a schematic (top) and as a raw alignment (bottom). The alignment is 69 residues in length, 66.7% of the residues are identical and has a statistically significant E()score of 2.4e-27. The black box indicates the boundaries of the alignment shown below. The colored boxes indicate Pfam domains, blue domains are SH3 domains, green are cortactin and pink are cofilin domains. An alignment calculated by the BLOSUM62 scoring matrix would be much longer.

information can only be transferred between those select residues.

Local alignments are also used for the more sensitive sequence comparison methods that use the transitive nature of homology or create position specific scoring matrices (PSSMs). The transitive inference of homology between sequences that are not statistically similar requires that the same region be shared across all of the sequences. PSSMs are built from substitutions seen within homologs from a single domain family. An initial pairwise similarity search is performed and the substitutions seen within inferred homologs are used to create a PSSM that is specific to that single domain. Next, that PSSM is used to search the database again and the results from that search are used to update the PSSM. This process of alignment and PSSM creation can be repeated as many times as necessary to identify more homologous sequences. Because PSSMs iteratively incorporate the substitutions seen in the alignments they create, PSSMs are very sensitive to alignment boundaries. Non-homologous residues that are included in alignments (at the boundaries) are then included in the PSSM and can lead to more erroneous alignments in further iterations (Gonzalez and Pearson, 2010a).

1.4 The Algebra of Similarity Scoring Matrices

The boundaries of a local alignment depend on the scoring matrix used to calculate the alignment. Local alignments require scoring matrices that have a negative average expected score; in other words, these matrices, on average, give alignments between randomly generated residues negative scores. If a scoring matrix with a positive average expected score were used, alignments would increase in score just by increasing in length, leading to a global alignment. In 1991, Altschul wrote a seminal paper outlining a universal way to evaluate scoring matrices (Altschul, 1991). He showed that every scoring matrix with a negative expected score behaves as a log-odds matrix. In amino acid substitution matrices the "odds" in the log-odds matrix is the ratio of the probability of a substitution occur-

ring as a result of evolution divided by the probability of the substitution happening by chance; that is, the score = $\log \frac{P(\text{Homology})}{P(\text{Chance})}$. Formally, this concept can be expressed as $S_{ij} = \log(\frac{q_{ij}}{p_i p_j}) / \lambda$, where the score (S_{ij}) given to a substitution is the frequency of the substitution from residue i to residue j seen in related sequences (q_{ij}) divided by the frequency of the residue i in all sequences (p_i) and the frequency of residue j in all sequences (p_j). Lambda is a scaling constant used to create scores that are in the same range (e.g., -10 – +15) independent of the matrix, and is part of the extreme value calculation.

1.5 Calculating Substitution Frequencies

Scoring matrices are calculated from a set of amino acid substitution frequencies (q_{ij}). The scoring matrices most widely used for protein alignment calculate substitution frequencies from a model of evolution (the PAM and VTML matrices) or measure them from direct observation of sets of related sequences (the BLOSUM matrices). The first model of evolution was based on point accepted mutations (PAMs) (Dayhoff *et al.*, 1978). PAMs are amino acid substitutions that were accepted by natural selection and became the dominant form of the protein sequence. In practice, PAMs are the substitutions that can be seen between related sequences.

The PAM model of evolution was created from a group of sequences that were at least 85% identical. Highly identical sequences were used to reduce the likelihood that an observed mutation (e.g., A-> D) is the result of multiple mutations to the same residue (e.g., A->?->D). The substitutions within these highly identical sequences were normalized to create a final substitution frequency matrix that reflected the expected substitution frequencies after 1 PAM. One PAM is equivalent to the evolutionary time it takes to acquire 1 accepted mutation per 100 residues or 99% identity between the sequences. As more evolutionary time passes, more PAMs are acquired and the identity between sequences reduces. The substitution frequencies for any distance of PAM (PAM(N)) can be calculated by mul-

tiplying the PAM1 substitution frequency matrix by itself N times.

Other scoring matrices updated the PAM model by re-calculating the PAM1 substitution frequency matrix using more data (Jones *et al.*, 1992; Gonnet *et al.*, 1992). In addition, the VTML series of matrices, while still based on a model of evolution, used new methodology to calculate the evolutionary model (Muller *et al.*, 2000, 2002). The PAM matrices and the derivatives of PAM relied on the PAM1 substitution frequency matrix to calculate a specific matrix. The VTML matrices do not rely on a single VTML1 substitution frequency matrix; instead they incorporate substitution data from sequences at many levels of divergence to estimate a substitution frequency matrix for any evolutionary distance. VTML matrices also use PAMs to denote evolutionary distance so the evolutionary distance reflected by VTML120 \sim PAM120. Using simulated alignments, the VTML model was shown to more accurately estimate substitutions seen across multiple PAM distances (Muller *et al.*, 2000, 2002).

The most popular scoring matrix used to create local alignments comes from the BLOSUM series of matrices (Henikoff and Henikoff, 1992). BLOSUM substitution frequencies are not derived from an evolutionary model but are measured directly from un-gapped blocks of related sequences at a specific identity cut offs. Changing the identity cutoff of the blocks included in the substitution frequency calculations changes the evolutionary distance that the matrix reflects. The substitution frequencies for BLOSUM62 were measured from blocks that were 62% identical or less, thus creating a matrix that is similar to PAM160 or VTML160, while BLOSUM80 was measured from blocks that were 80% or less identical, thus leading to a matrix that is similar to PAM40. BLOSUM62, the matrix most often used for inferring homology, has been shown to be very sensitive. This particular matrix is able to find more homologs than many of the PAM matrices (Henikoff and Henikoff, 1992) as well as the updated version of the PAM matrices that were derived from more data (Henikoff and Henikoff, 1993). In a direct comparison between BLOSUM62 and the evolutionarily similar VTML160, BLOSUM62 was able to identify more homologs in

108 of the 136 protein families queried (data collected by author), once again establishing BLOSUM62's improved sensitivity over other matrices.

1.6 Scoring Matrix Depth and Alignment Length

Different substitution frequencies in scoring matrices lead to different alignments. Deep scoring matrices are designed to reflect substitutions frequencies over long evolutionary distances while shallow matrices reflect substitution frequencies after short evolutionary distances. Deep matrices are most often used during similarity searches because they are the most sensitive and are able to identify very distant homologs. Figure 1.2A shows the substitution frequencies for VTML120 and VTML20. VTML20 is a very shallow scoring matrix representing the substitutions seen between sequences that are 85% identical; VTML120 is 6 times deeper than VTML20 and represents sequences that would be about 40% identical. The differences in identity between VTML20 and VTML120 are reflected in their substitution frequencies. The frequencies for identical substitutions (along the diagonal) for VTML20 are very close to 1 (0.82 for A:A) while the substitutions for the same identities in the VTML120 matrix are less than half as frequent (0.33 for A:A). The opposite trend is seen for non-identical substitutions; VTML20 substitutions between non-identities are very low (0.008 A:R) while VTML120 has much higher non-identical substitution frequencies (0.04 A:R). Differences in substitution frequencies are carried over into the final scoring matrix as differences in final scores (Figure 1.2B). The higher identity substitution frequencies seen in VTML20 result in larger positive scores given to identities (7 for A:A) compared to VTML120 (4 for A:A). Higher non-identity substitution frequencies seen in VTML120 result in less negative (-2 for A:R) and sometimes positive scores (2 for N:D) given to non-identical substitutions compared to VTML20 (-7 for A:R, -1 for N:D). Differences in amino acid abundances ($p_i p_j$) also affect the final score for a substitution, though these values are constant in every matrix and do not account for either the differences in

scores or alignments between different matrices. The differences in the magnitude of both positive and negative scores of each matrix lead to the differences in alignments created by each matrix.

Once Altschul described scoring matrices as log-odds matrices, he was able to apply principles from information theory to describe each matrix (Altschul, 1991). Entropy (H), the average information content per aligned residue, can be used to describe the relative magnitude of the scores between two matrices. Matrices with higher entropy will have larger magnitude scores (both positive and negative) than a matrix with lower entropy. Deep scoring matrices have low entropy; each aligned residue can only help or hurt the overall similarity score a little. Shallow matrices have much higher entropies; each aligned residue has a large impact on the overall similarity score. VTML120 (entropy 0.94 bits per aligned residue) requires 3 times the number of aligned residues to gain or lose the same magnitude of score as it would for VTML20 (entropy 2.96 bits per aligned residue). The effect of different scoring matrices on alignments is shown in Figure 1.2C. Alignments and scores between the same sequences are shown using three different scoring matrices, VTML20 is the shallowest, VTML120 is intermediate and BLOSUM62 is the deepest, as denoted by the entropy of each matrix (Figure 1.2C column H). The alignment (red boxes) created by VTML20 is much shorter than the alignments created by both VTML120 and BLOSUM62. The highest cumulative score for the VTML20 alignment is reached after the first two residues are aligned; that is, the magnitude of the negative scores could not be overcome by the additional C:C identity. This is not true for the alignments created by BLOSUM62 and VTML120, where the magnitude of the negative scores were not as large and the additional C:C identity brought the cumulative score above (BLOSUM) or equal to (VTML120) the previous maximum cumulative score, thus allowing the alignment to continue. Scoring matrix effects can also be seen in the cumulative score plot shown for the alignment between Human cortactin and Xenopus dbnl-b using BLOSUM62 and VTML40 (Figure 1.3, alignment also shown in Figure 1.1B). In the alignment scoring

A) Substitution Frequency Matrix

VTML120						VTML20					
	A	R	N	D	C		A	R	N	D	C
A	0.3277					A	0.8157				
R	0.0391	0.3969				R	0.0080	0.8442			
N	0.0488	0.0399	0.3053			N	0.0096	0.0092	0.80982		
D	0.0466	0.0169	0.0818	0.3925		D	0.0101	0.0007	0.02952	0.8442	
C	0.0919	0.0199	0.0180	0.0096	0.4558	C	0.0273	0.0043	0.00319	0.0001	0.8757

B) Final Scoring Matrix

VTML120						VTML20					
	A	R	N	D	C		A	R	N	D	C
A	4					A	7				
R	-2	6				R	-7	8			
N	-1	-1	6			N	-6	-5	8		
D	-1	-3	2	6		D	-6	-12	-1	8	
C	0	-3	-3	-5	9	C	-3	-7	-8	-14	11

C) Local Alignment

		seq1		N	R	R	D	N	C
		seq2		N	R	D	C	R	C
H									
0.58	BLOSUM62		6	5	-2	-3	0	9	
				11					15
0.94	VTML120		6	6	-3	-5	-1	9	
				12					12
2.67	VTML20		8	8		-12	-14	-5	11
				16					-3

Figure 1.2: Scoring matrices at different evolutionary distances. A) The substitution frequencies are given for a 5x5 section of both VTML120 and VTML20. Below, the final scores corresponding to the substitution frequencies above are shown. B) An alignment is calculated using BLOSUM62 (BL62), VTML120 and VTML20. The entropies for each of the scoring matrices is given by H. The boundaries of the final alignment are denoted by the red box. Cumulative scores are given by the subscript with the maximum cumulative score denoted in red.

by both VTML40 and BLOSUM62 most of the score is accumulated from the region of the alignment that corresponds to the SH3 domain (Figure 1.3 top). The region outside of the SH3 domain are scored very differently by the two matrices. VTML40 penalizes the residues outside of the SH3 domain very harshly while BLOSUM62 has a slightly positive score over this region. The BLOSUM62 alignment is 200 residues long while the VTML40 alignment is only 60 residues long and more closely matches the location of the SH3 domain.

1.7 Homologous Over Extension

This dissertation examines alignment boundary accuracy. Homologous sequences have the same structure, do when regions of sequences are identified as homologous, the boundaries of the alignment indicate the boundaries of the structural similarity. To maximize structural similarity, alignments should only contain residues that exist in a single structural unit (e.g., alpha-helix or beta-pleated sheet). Alignments with accurate boundaries exclude non-homologous residues and include all possible homologous residues.

Alignment boundary accuracy is different from equivalent residue accuracy measured in multiple sequence alignment (MSA) (Aniba *et al.*, 2010; Thompson *et al.*, 2011). A MSA can be thought of as a continuation of multiple pairwise alignments; once many homologs have been identified using pairwise methods, all of the homologous sequences can then be aligned with the aim to identify equivalent residues across all of the homologs. In practice, accurately identifying equivalent residues in a MSA relies on the accurate placement of gaps to account for insertions and deletions that have occurred throughout the sequences evolutionary history. Accurate versus inaccurate MSAs only differ by where the gaps are placed; it is assumed that an MSA represents a single structural unit.

By recognizing alignments that have included non-homologous residues, we can establish an alternative hypothesis when structural and sequence alignments do not match.

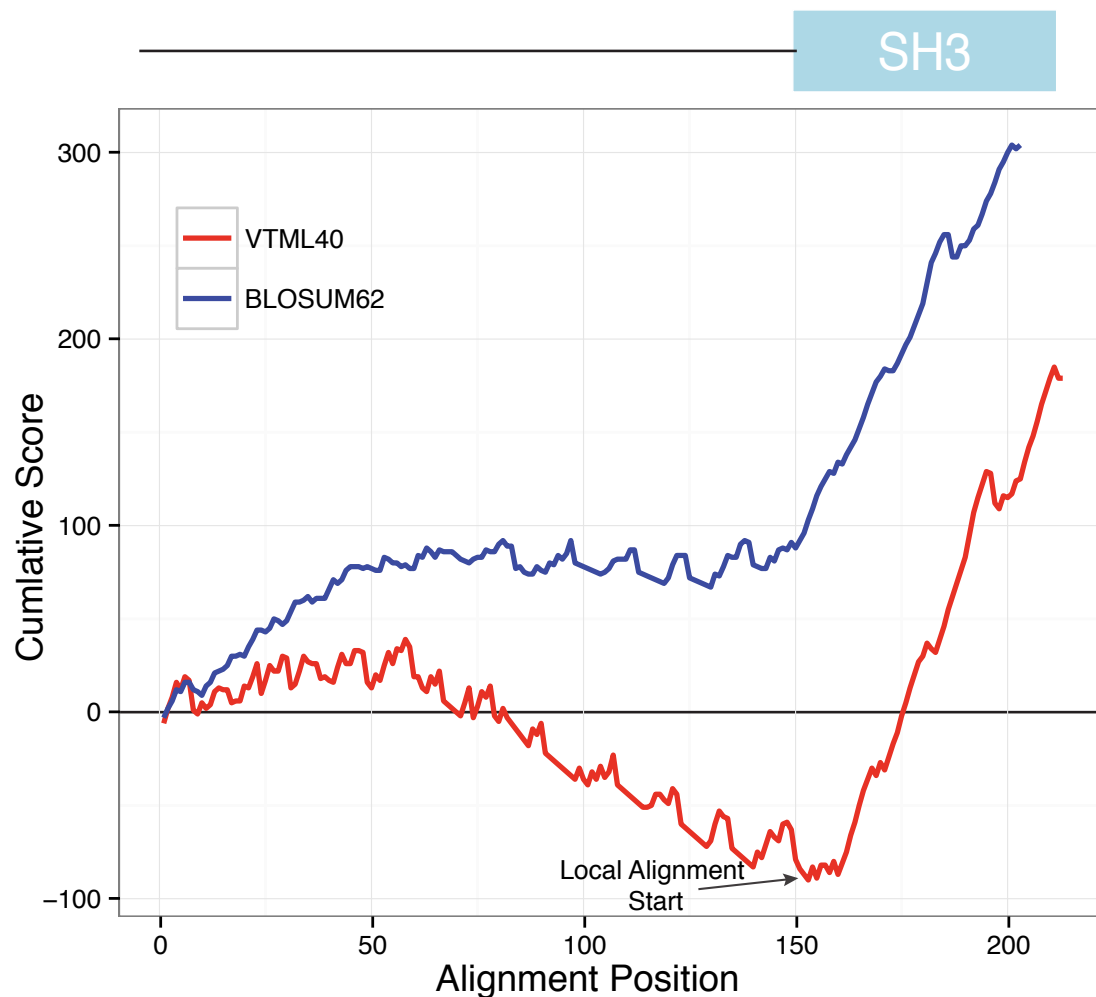


Figure 1.3: Cumulative score plots within BLOSUM62 local alignment boundaries. An alignment between SRC8_HUMAN residues 350-330 and DBNLB_XENLA residues 195-375 was scored using both BLOSUM62 (blue) and VTML40 (red). The protein coordinates used reflect local alignment boundaries created by BLOSUM62 between the same sequences. The beginning of the local alignment created by VTML40 between the sequences is denoted by the arrow. A schematic of the location of the SH3 domain in the alignment is given at the top.

Homologous over extension (HOE) was first identified as a source of error during iterative similarity searches (Gonzalez and Pearson, 2010a). HOE occurs when alignments between homologous sequences extend beyond the boundaries of the homologous region and include non-homologous residues from either of the aligned sequences. The addition of non-homologous residues was found to corrupt the position specific scoring matrices that were built by the iterative search algorithms, thus leading to the identification of sequences that were homologous to the additional non-homologous residues and not the domain originally used as a query.

Roessler et al. published an alignment between two Cro family members from two bacteria phages P22 and λ . The structures for both of the proteins are known, the protein from P22 only contains alpha-helices while the protein from λ contains both alpha-helices and beta-pleated sheets. The 65-residue alignment was 40% identical and achieved a statically significant level of similarity, but this alignment also aligned residues involved in the beta-sheet of the λ protein to the alpha-helices in the protein from P22. The authors used this sequence alignment as evidence that sequences can be similar without implying similar structures. Now that some evidence exists that alignments do not always have accurate boundaries, perhaps the anomaly seen in the alignment between these two bacteria phage proteins could more easily be explained by inaccurate alignment boundaries and not as the first example of an otherwise unseen evolutionary jump between two different protein structures.

Possible HOE has also been observed in DNA alignments between highly conserved, evolutionarily close genomes (e.g.; *C. elegans* and *C. briggsae*) (Arslan *et al.*, 2001; Chao *et al.*, 1993; Zhang *et al.*, 1999). Here, genomic alignments included very low identity regions, usually corresponding to introns and intergenic regions, dispersed between very highly conserved sequences corresponding to exons or whole genes. While these long alignments had statistically significant similarity scores and were homologous, it was obvious from closer inspection that the alignment, while optimal in the sense that it was the

highest score possible between those two sequences, did not accurately reflect the underlying biology (exon/intron structure) of the aligned genomic regions.

Accurate alignment boundaries impact homology inference; alignment boundaries also impact the downstream applications that rely on an initial alignment. Local alignments are the basis for iterative search strategies that work to identify more distant homologs as well as for multiple sequence alignment, which are then the basis for building phylogenetic trees. Local alignment methods are also used in the genomics field to identify structural variants and breakpoint locations. While the ability of similarity searching to identify homologous sequences is not in question the alignment boundaries, even between extremely significantly similar sequences, warrants examination. The boundaries of a local alignment depend on the scoring matrix used to calculate the alignment. Each scoring matrix has a specific evolutionary distance that it is designed to reflect and the evolutionary distance of a scoring matrix changes the alignments that the matrix creates. In chapter 2 of this dissertation I will show work testing the hypothesis that non-homologous residues are included in alignments because of a mismatch between the evolutionary distance of the sequences and the evolutionary distance of the scoring matrix used to create the alignment.

Chapter 2

2.1 Abstract

2.1.1 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.

2.1.2 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.

2.2 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. 2.1). 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. 2.1 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 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.

2.3 Methods

2.3.1 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 (pres-

ence 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.

2.3.2 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 2.1.

2.3.3 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.

2.3.4 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

Table 2.1: Scoring matrices, gap penalties, and mean identity, entropy, and alignment length. *Means measured from 136 random sequence searches (Fig. 2.3).

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

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.

2.3.5 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 2.2.

2.4 Results

2.4.1 Homologous over-extension

Deep scoring matrices can produce inaccurate alignment boundaries. Fig. 2.1 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, 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.

2.4.2 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. 2.2A). 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

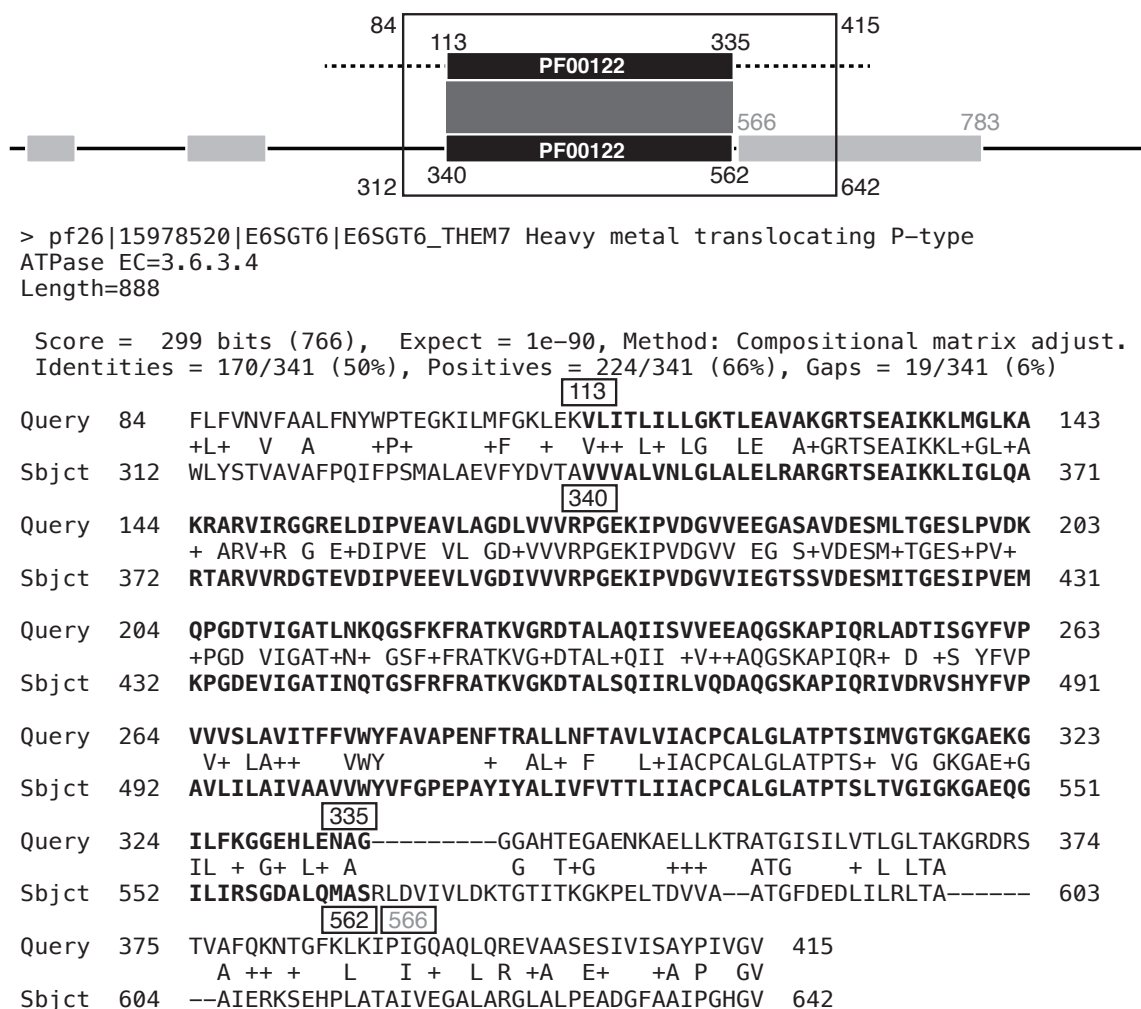


Figure 2.1: **Homologous overextension.** BLASTP with BLOSUM62 was used to create an alignment between a RPD2 query and a homologous sequence from the RPD2 library. The raw BLASTP output and a schematic of the sequences are shown. Homologous domains in the query (top) and subject (bottom) sequence are 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.

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. 2.2B).

Fig. 2.2 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.

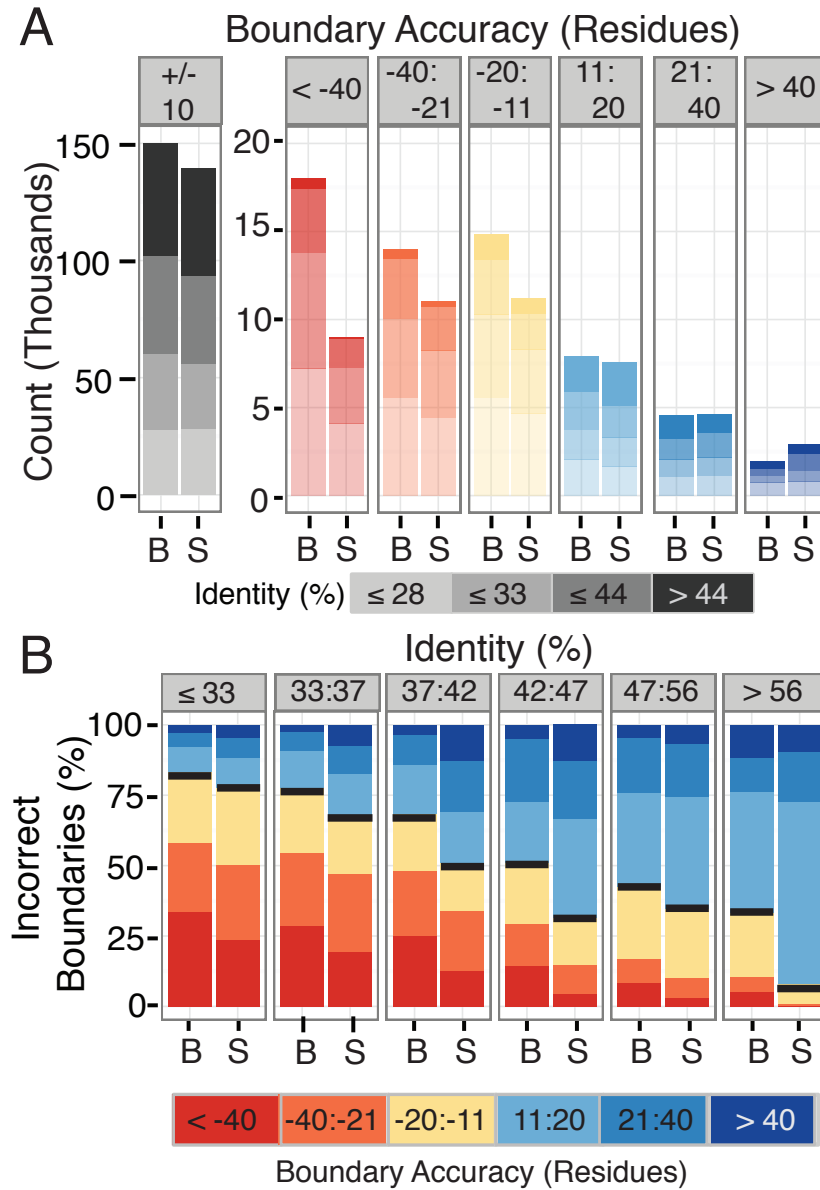


Figure 2.2: **Boundary accuracy and sequence identity.** Using the RPD2 embedded domain queries and sequence library, pairwise protein sequence alignments were calculated with BLASTP (B) and SSEARCH (S) using BLOSUM62. Boundary accuracy was measured for both the beginning and end of alignments between 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.

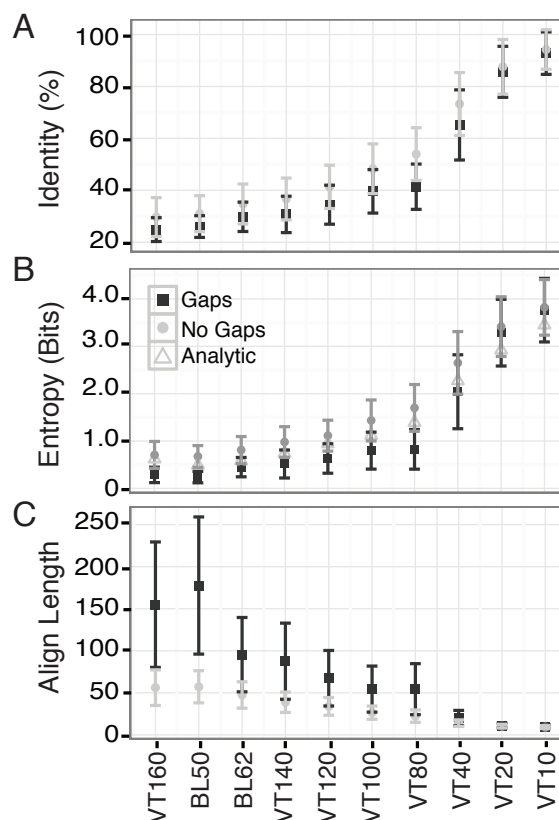


Figure 2.3: **Scoring matrix target identity, entropy, and alignment length.** Queries were constructed from 136 shuffled protein domains. SSEARCH was used to search against the RPD2 library with these shuffled queries using either the gap penalties given in Table 2.1 (black squares) or gap penalties 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).

2.4.3 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. 2.3). 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.

2.4.4 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 with 16,640 over-extended alignments, we tracked the boundary accuracy produced by six VTML matrices (VT) with increasing target identity (Fig. 2.4). The alignment with the smallest cu-

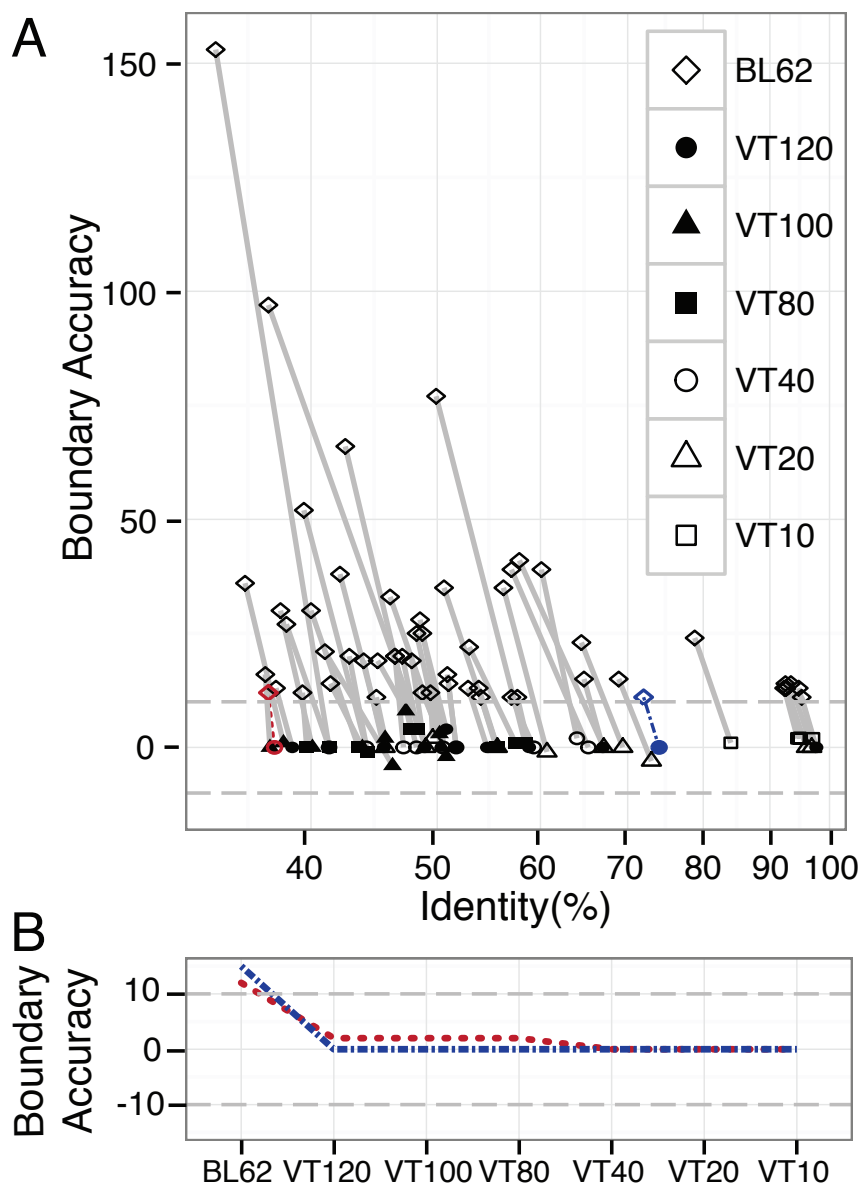


Figure 2.4: **Selecting the scoring matrix that creates the best alignment.** (A) Sequence pairs with $>33\%$ identity and over-extended alignment boundaries were selected from the results of the similarity search performed using SSEARCH with BLOSUM62. Each sequence pair was realigned using VT120,100,80,40,20 and VT10 (Table 2.1). 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) shows a low identity alignment corrected by VT40; the square dash-dot line to the right in panel (A) shows a high-identity alignment corrected by VT120.

mulative 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. 2.4. 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 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. 2.4B 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. 2.4),

we attempted to correct BLOSUM62 alignments using the scoring matrices and thresholds shown in Table 2.2.

Table 2.2: Identity required to re-align using each scoring matrix. Values are inclusive at the maximum for each matrix.

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

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 2.2 dramatically decrease over-extension errors they can also produce incomplete alignments (Fig. 2.5). In contrast to Fig. 2.4, where we selected the most accurate alignment, Fig. 2.5 shows the results of re-alignment based solely on the identity of the initial BLOSUM62 alignment (the thresholds in Table 2.2). 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. 2.2) 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.

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.

2.5 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. 2.1, Fig. 2.2). 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. 2.3). 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. 2.2). Using a shallower scoring matrix that targets the correct sequence identity can correct over-extension (Fig. 2.4). 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 be-

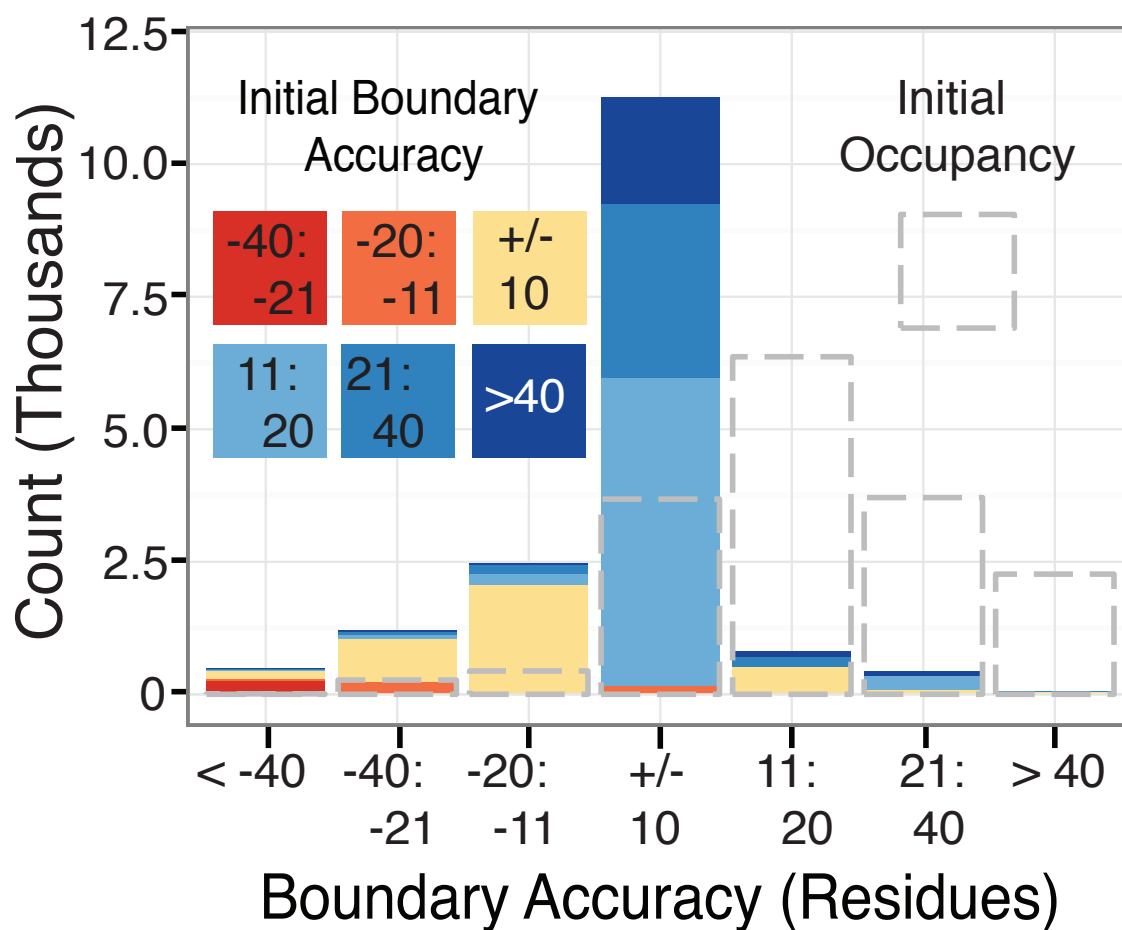


Figure 2.5: **Re-Alignment Algorithm Results** Only results from sequence pairs that were re-aligned by the algorithm are shown. Grey dashed bars indicate the boundaries in each accuracy bin before re-alignment; colored bars indicate the final distribution of alignment boundary errors. The colors in the final distribution bars show the original accuracy before re-alignment.

tween 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. 2.3). 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. 2.3). 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. 2.1, 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 di-

verse. 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 $2 \times 200,000$ 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 mul-

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, this 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. Surprisingly, 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. 2.3). The long alignments in Fig. 2.3 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 occurred 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 align-

ment, 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.

Chapter 3

3.1 Summary of Major Conclusions

In the past, similarity search methods have been evaluated based on sensitivity; algorithms were only judged based on how many known homologs they could detect. This focus on sensitivity produced methods that are able to identify extremely distant homologs (Henikoff and Henikoff, 1992; Brenner *et al.*, 1998; Henikoff and Henikoff, 1993; Pearson, 1995). Studies that evaluate sensitivity have not examined the accuracy of the alignments that identified each homolog. This focus on sensitivity creates a problem; the most sensitive methods do not always create accurate alignment boundaries. Disagreements between structural and sequence alignments, as well as excessively long genomic alignments that include regions of the genome not thought to be highly conserved (e.g., introns), suggest that inaccuracies can exist in local alignment boundaries.

Building on work that first identified the homologous over extension (HOE) during iterative search strategies, here we have measured the frequency (8%) and extent to which alignment boundary inaccuracy occurs during typical pairwise similarity searches (BLAST searches). HOE was seen at all identity levels, but the frequency of HOE was positively correlated with the identity of the aligned sequences. In fact, in this study identity was the only reliable predictor of HOE. In contrast to search sensitivity, where the most distantly related homologous are missed, HOE occurs more frequently in closely related homologs.

The frequency of HOE increased to 10% in the top 100 scoring alignments as compared to all alignments that reach statistical significance. So, while HOE is an error, it is not in the ability to identify homologous sequences, but it is an inability to place the correct boundaries on the homologous region.

Local alignment boundaries depend on the scoring matrix used to calculate the alignment. Altschul provided a theoretical framework (entropy) to explain how scoring matrices designed for different evolutionary distances affect alignments (Altschul, 1991). However, his calculations only applied to local alignments that did not include gaps. We have extended his mathematical observations by directly measuring a scoring matrices target identity, entropy and expected alignment lengths in alignments that include gaps. These intrinsic properties of scoring matrices, specifically the changes in alignment length, provides both an explanation for HOE, and a strategy for correcting it. This work also links the abstract notion of evolutionary distance of a scoring matrix to a concrete and simple to measure variable: sequence identity. Using sequence identity to match specific alignments to the correct scoring matrix can correct alignment boundaries, thus removing non-homologous residues from alignments. The ability to correct alignments using sequence identity strengthens the link between the evolutionary distance between two sequences, the evolutionary distance of the scoring matrix, and the accuracy of alignment boundaries.

3.2 Other Algorithms to Identify Non-Homologous Residues

Matching the evolutionary distance of the alignment with the correct scoring matrix results in more accurate alignment boundaries, but a more universal method to evaluate alignment boundary accuracy is still needed. Alignments with over extended boundaries provide underestimates of the correct alignment's identity and therefore evolutionary distance. While this inaccurate identity measurement can be used to select the correct scoring matrix and correct many alignments, some alignments were still over extended, while others became

incomplete. There were also many over extended alignments with identities that fell below our threshold for switching scoring matrices (36%), and therefore never had the opportunity to be corrected.

Every over extended alignment will have the greatest identity and accumulation of similarity score focused in the homologous domain, while the excess residues should have much lower identity and very little score accumulation (Figure 1.3). Currently, a new scoring matrix can only be selected if the identity of the truly homologous domain of the alignment is high enough that the additional low identity over extension does not result in an alignment that has less than 36% identity. Peak finding algorithms (Rashid *et al.*, 2011; Zhang *et al.*, 2008) can identify the higher scoring, higher identity, sub-regions of an alignment from the lower identity, lower scoring over extended sub-region(s). By scanning along an alignment, peak finding algorithms may identify sub-regions of the alignment that contain higher mean scores than alignment as a whole thereby separating the truly homologous domain from the excess non-homologous residues. Alignments with correct boundaries, or those with incomplete alignments may not have a region of higher identity or higher score within them. But, for over extended alignments where a sub-region of higher score or identity was found, the boundaries of the sub-region could be more accurate boundaries for the homologous domain than the original boundaries created by the alignment.

Another method that could be used to identify a sub-region of higher scoring, higher identity residues within an alignment, regardless of the alignments identity, is change-point detection. Change-point detection algorithms are used to identify changes in the mean of a process (Assareh *et al.*, 2011; Muggeo and Adelfio, 2011) . This methodology could be used to detect shifts in the score or identity of an alignment as it is accumulated across the alignment. A shift between a lower scoring, lower identity sub-region into a higher scoring, higher identity one would indicated the beginning of the homologous domain (like the shift between the non-domain to domain region in Figure 1.3) while a shift from a higher scoring, higher identity region to a lower scoring, lower identity region would indicate the end of the

homologous domain. As with peak finding a lack of change-point detection would indicate accurate domain boundaries or an incomplete alignment.

3.3 Sub-Alignment Scoring and Incorporation of Outside Annotations

While peak finding and change point detection algorithms could be used to detect shifts in scores from the alignment alone, leveraging domain annotation information could also lead to more accurate alignment boundaries. Peak finding and change-point detection seek to subdivide an alignment into multiple parts. For example, in Figure 1.3 the over extended part of the alignment has a slowly increasing score with BLOSUM62, while the score increases rapidly when the homologous SH3 domains are aligned. In contrast to peak finding, we can use external domain annotation to subdivide the alignment and the score. The current version of the FASTA programs allows alignments to be subdivided based on residue coordinates provided by Uniprot, Interpro or Pfam. A raw alignment score and percent identity is calculated within a sub-alignment; the raw score is also normalized to a bit-score and converted to an sub-alignment score probability.

Using domain information from databases such as Pfam (Punta *et al.*, 2012) or Interpro (Hunter *et al.*, 2012) to annotate sequences allows scores to be associated with specific domains. HOE could be inferred when two conditions are met: (1) when the alignment boundaries extend beyond the annotated domain boundaries and (2) when the non-domain region is much lower scoring (even for its length) than the flanking domain annotated region. Unlike peak finding or change point detection, sub-alignment scoring can distinguish between correct alignment boundaries and incomplete alignments. Alignments that begin and end within annotated domain boundaries are incomplete and require more sensitive methods to detect the full length of the homologous domain, while correct alignment boundaries will match the annotation.

3.4 Complex Protein Architectures

We believe this work underestimates the amount of HOE that occurs during similarity searches. Alignment boundary accuracy was measured using a simplified version of protein architecture; each query sequence only contained a single domain and each of the domains were longer (>200 residues) than the average alignment BLOSUM62 would be expected to create by chance (150 residues). Many proteins have multiple domains, 40% of the human sequences in Pfam have more than one domain and many common domains like SH3 and Ankyrin repeats are much shorter than 150 residues. While a strong correlation between domain length and the amount of over-extension was not found in long domains, it may emerge when domain lengths are below BLOSUM62's average expected alignment length.

Many proteins contain multiple domains and multi-domain structures are shared across homologs. When two different protein sequences share multiple domains does over-extension increase in frequency or extent? If the domains are separated by a non-domain region, how can we know whether the non-domain region should be included in the alignment or is just a product of over-extension? Experiments using sub-alignment scoring could lay the groundwork to understanding the scores that are seen in sub-alignments that contain known non-homologous (e.g., random) sequence. This distribution of scores could then be used as a guideline for minimum total score, bits per position, identity or statistical cutoffs to begin to delineate between over-extension and excess similarity that exists outside of domain boundaries.

3.5 Boundary Accuracy in DNA Alignments

The scoring matrices used to align genomic DNA sequences also represent a specific evolutionary distance that can be expressed in PAMs and target identity. Like scoring matrices used for protein sequence alignment, each DNA scoring matrix should have a range of PAM distances and optimal sequence identity range. The effective PAM range of several DNA

matrices was analytically derived using entropy calculations from DNA matrices designed for different PAM distances (States and Botstein, 1991). Like the data shown for amino acid scoring matrices (in my work and Altschul (1991)), each matrix had a unique range of efficiency and shallow DNA matrices were the most efficient at short PAM distances while deeper DNA matrices were the most efficient at longer PAM distances.

Expected conservation between different genomes is location dependent; exons are more similar across species than introns or intergenic regions. Some special regions of the genome, promoters, 5' and 3' UTRs, intron splice donor and acceptor sites as well as enhancer regions may share an intermediate level of similarity to that of exons and introns. Human and mouse exons are on average 85% identical, while introns between the species share an average of 35% identity at the DNA level and do not share statistically significant similarity. This large disparity in expected identity is not taken into account when a single scoring matrix is used to align closely related genomes. The +1 for a match, -1 for a mismatch, scoring matrix often used to align highly related genomes is equivalent to 30 PAM or 75% identity. The expected identity from this matrix is in between the expected identity of exons and introns, and may not be ideal for either. Using a shallower matrix designed to have a target identity that is correct for exons would limit HOE but it would also limit the ability to detect homology that may exist between less conserved regions. Sub-alignment scoring could allow for the use of different matrices to align different genomic regions depending on expected identity levels. This would allow for the recognition of conserved regions in lower identity regions of the genome without the influence of high identity neighboring sequences.

With the advent of next-generation sequencing, new alignment tools were needed to align millions of sequences (millions of queries) to a single reference genome (database of one long sequence). This process is impractically slow using BLAST or FASTA like aligners that are designed to query a large database with a single sequence. The methods developed to make next-generation aligners computationally efficient depend on the sequenced

genome being highly identical ($> 97\%$) to the reference. Natural genome variation and the error rates of next-generation sequencers both contribute to the sequenced genome's divergence from the reference. In addition to genomic divergence, the background error rate of next-generation sequencers adds additional "divergence time" to the natural genomic variation that researchers often want to capture when sequencing a genome.

When next-generation sequence aligners first came out, they were working with sequences (reads) that were less than 100 nucleotides, and created global alignments between the reads and the reference genome (Li and Durbin, 2009). With sequences this short, SNPs, insertions and deletions were rare, so the reads were highly identical to the reference. High identity makes global alignments easy to find, makes identification of structural variation, with its local disruption of alignment boundaries, more difficult. With increases in read length to over 100 nucleotides, structural variations can be identified, but reads can also become more divergent from the reference so global alignments are less efficient. Long-read aligners can create local alignments between sub-regions of reads and the reference genome (Li and Durbin, 2010). But these aligners still only tolerate low levels (2% BWA-SW) of divergence between the reads and the reference within the local alignment. In studies to identify structural variations in the genome we expect that localized regions of the sequenced genome will be highly divergent from the reference and therefore hard to align with confidence using aligners that can only tolerate low levels of divergence. The regions of the genome that are involved in structural variation are still a small minority compared to the regions of the genome that differ from the reference by sequencing errors, so allowing more divergence across every read could lead to genome wide alignment inaccuracies. Current long read aligners used to detect structural variation use a single scoring matrix targeted to the expected error rate of the sequencer to create alignments. Targeting reads or groups of reads with different scoring parameters may lead to a greater number of reads being aligned at all, and a better estimation of where structural variations are occurring.

3.6 Improving Iterative Search Strategies

HOE was first identified as causing errors in PSI-BLAST searches leading to known non-homologs achieving statistically significant levels of similarity (Gonzalez and Pearson, 2010a). During iterative searches, HOE has its largest effects in the later PSSM iterations. PSSMs do not have target identities and cannot be matched to the evolutionary distance between two sequences, so the methods I describe for pairwise alignment cannot easily be extended to PSI-BLAST or PSI-SEARCH. A PSSM built from a given initial query sequence will be the full length of the entire query. If the local pairwise alignments only encompass part of the full length query, then the data in the PSSM from that region will come directly from the alignments while the data outside of this region will look like a deep scoring matrix (e.g, BLOSUM62 for PSI-BLAST). When the PSSM is used to search the database during the next iteration, it will give very high scores to the same sequences used to build the PSSM specifically across the homologous regions. The regions outside of this homologous region will be scored similarly to a deep scoring matrix, which we now understand do not assign large enough penalties to mismatches allowing the alignment to extend past the boundaries of the homologous domain.

Limiting HOE by not allowing sequences that have already been aligned once to grow in length during later search iterations can improve PSI-BLAST performance by limiting over extension. But, limiting extension also makes searches less sensitive (Li *et al.*, 2012). Sub-alignment scoring might be used to limit HOE while maintaining sensitivity. When an alignment between the PSSM and a sequence expands during a later iteration, sub-alignment scoring could be used to independently evaluate the additional residues to determine if they are included because of HOE or enhanced domain detection. The residues included because of HOE could be excluded from the PSSM, reducing errors. Incomplete alignments that expand during further iterations would be allowed to expand and included in the PSSM, perhaps leading to a boost in the PSSM sensitivity.

3.7 Final Thoughts

Understanding the possible errors that could lead to false positives or false negatives is key to understanding data from any computational tool. Homologous over extension is a new form of false positive that occurs during similarity searches and is caused by errors in alignment. Local alignment boundaries are dependent on the scoring matrix used to calculate the alignment. This work increased our understanding of the properties of the alignments created by different scoring matrices by measuring them directly. This work provides more insight into the interaction between the evolutionary distance of the scoring matrix, the evolutionary distance of the homologous residues and the alignment that was calculated leading to a better understand of when and why alignment boundary inaccuracies occur.

Alignment is a key component of several computational methods used to explore a wide range of biological questions, and each of these methodologies has their own strengths and weaknesses. Better understanding of the errors associated with alignment, including alignment boundary errors, may lead to a greater understanding of the other issues seen in computational methods that use alignment. This in turn could lead to more accurate computational tools and a better understanding of the underlying biology explored using these tools.

Bibliography

Altschul, S., Boguski, M., Gish, W., and Wootton, J. (1994). Issues in searching molecular sequence databases. *Nature Genetics*, **6**, 119–129.

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

Altschul, S. F. and Gish, W. (1996). [27] local alignment statistics. In R. F. Doolittle, editor, *Computer Methods for Macromolecular Sequence Analysis*, volume 266 of *Methods in Enzymology*, pages 460 – 480. Academic Press.

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.

Aniba, M. R., Poch, O., and Thompson, J. D. (2010). Issues in bioinformatics benchmarking: the case study of multiple sequence alignment. *Nucleic Acids Research*, **38**(21), 7353–7363.

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

Assareh, H., Smith, I., and Mengersen, K. (2011). Change point detection in risk adjusted control charts. *Statistical Methods in Medical Research*.

- 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.
- Doolittle, R., Hunkapiller, M., Hood, L., Devare, S., Robbins, K., Aaronson, S., and Antoniades, H. (1983). Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor. *Science*, **221**(4607), 275–277.
- Eddy, S. R. (2011). Accelerated profile hmm searches. *PLoS Comput Biol*, **7**(10), e1002195.
- Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**(5), 1792–1797.
- Feng, D., Johnson, M., and Doolittle, R. (1985). Aligning amino acid sequences: comparison of commonly used methods. *Journal of Molecular Evolution*, **21**, 112–125.
- Fitch, W. (1976). The molecular evolution of cytochrome c in eukaryotes. *Journal of Molecular Evolution*, **8**(1), 13–40.
- Gonnet, G., Cohen, M., and Benner, S. (1992). Exhaustive matching of the entire protein sequences database. *Science*, **256**(5062), 1443–1445.

- 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 (Oxford, England)*, **26**(18), 2361–2362. PMID: 20693322.
- Haber, J. E. and Jr., D. K. (1970). An evaluation of the relatedness of proteins based on comparison of amino acid sequences. *Journal of Molecular Biology*, **50**(3), 617 – 639.
- Henikoff, S. and Henikoff, J. (1992). Amino-acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, **89**, 10915–10919.
- Henikoff, S. and Henikoff, J. G. (1993). Performance evaluation of amino acid substitution matrices. *Proteins*, **17**, 49–61.
- Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., de Castro, E., Coggill, P., Corbett, M., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Fraser, M., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., McMenamin, C., Mi, H., Mutowo-Muellenet, P., Mulder, N., Natale, D., Orengo, C., Pesseat, S., Punta, M., Quinn, A. F., Rivoire, C., Sangrador-Vegas, A., Selengut, J. D., Sigrist, C. J. A., Scheremetjew, M., Tate, J., Thimmajananathan, M., Thomas, P. D., Wu, C. H., Yeats, C., and Yong, S.-Y. (2012). Interpro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Research*, **40**(D1), D306–D312.
- Jones, D., Taylor, W., and Thornton, J. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, **8**(3), 275–282.
- Karlin, S. and Altschul, S. F. (1990). Methods for assessing the statistical significance

- of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences*, **87**(6), 2264–2268.
- Karlin, S. and Altschul, S. F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Sciences*, **90**(12), 5873–5877.
- Koestler, T., von Haeseler, A., and Ebersberger, I. (2012). REvolver: modeling sequence evolution under domain constraints. *Molecular Biology and Evolution*, **29**(9), 2133–2145.
- Koonin, E. and Galperin, M. (2003). Chapter 2: Evolutionary concept in genetics and genomics. In *Sequence-Evolution-Function: Computational Approaches in Comparative Genomics*. Kluwer Academic.
- Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with burrowsDwheeler transform. *Bioinformatics*, **25**(14), 1754–1760.
- Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with burrowsDwheeler transform. *Bioinformatics*, **26**(5), 589–595.
- Li, W., McWilliam, H., Goujon, M., Cowley, A., Lopez, R., and Pearson, W. R. (2012). PSI-Search: iterative HOE-reduced profile SSEARCH searching. *Bioinformatics (Oxford, England)*, **28**(12), 1650–1651. PMID: 22539666.
- McLachlan, A. (1971). Tests for comparing related amino-acid sequences. cytochrome c and cytochrome c551. *Journal of Molecular Biology*, **61**(2), 409 – 424.
- Muggeo, V. M. R. and Adelfio, G. (2011). Efficient change point detection for genomic sequences of continuous measurements. *Bioinformatics*, **27**(2), 161–166.
- Muller, T., Spang, R., and Vingron, M. (2000). Modeling amino acid replacement. *Journal Comp. Bio.*, **7**(6), 761–776.

- 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.
- Needleman, S. and Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, **48**(3).
- Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases. *Protein Sci.*, **4**, 1145–1160.
- Pearson, W. R. (1996). [15] effective protein sequence comparison. In R. F. Doolittle, editor, *Computer Methods for Macromolecular Sequence Analysis*, volume 266 of *Methods in Enzymology*, pages 227 – 258. Academic Press.
- 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.
- Rashid, N. U., Giresi, P. G., Ibrahim, J. G., Sun, W., and Lieb, J. D. (2011). ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biology*, **12**(7), R67. PMID: 21787385.

- Reese, J. T. and Pearson, W. R. (2002). Empirical determination of effective gap penalties for sequence comparison. *Bioinformatics*, **18**, 1500–1507.
- Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, **147**(1), 195–197. PMID: 7265238.
- States, D. J. and Botstein, D. (1991). Molecular sequence accuracy and the analysis of protein coding regions. *Proceedings of the National Academy of Sciences*, **88**(13), 5518–5522.
- Thompson, J. D., Linard, B., Lecompte, O., and Poch, O. (2011). A comprehensive benchmark study of multiple sequence alignment methods: Current challenges and future perspectives. *PLoS ONE*, **6**(3), e18093.
- Wagner, G. P. (1989). The origin of morphological characters and the biological basis of homology. *Evolution*, **43**(6), pp. 1157–1171.
- Zhang, Y., Liu, T., Meyer, C., Eeckhoutte, J., Johnson, D., Bernstein, B., Nusbaum, C., Myers, R., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of chip-seq (macs). *Genome Biology*, **9**(9), R137.
- Zhang, Z., Berman, P., Wiehe, T., and Miller, W. (1999). Post-processing long pairwise alignments. *Bioinformatics*, **15**(12), 1012–1019.