WAS IT A MATch I SAW? Approximate palindromes lead to overstated false match rates in benchmarks using reversed sequences

George Glidden-Handgis[1] **and Travis J. Wheeler**[1]

[1]R. Ken Coit College of Pharmacy, University of Arizona

Accurate labeling of biological sequences requires an effective mechanism for evaluating the rate of labeling error. In the context of sequence annotation, tool developers perform this evaluation using benchmarks that consist of trusted homologs for measuring sensitivity and some form of decoy sequences for measuring false labeling. Various methods exist for producing decoys, including shuffling and random sampling; however, these methods may underestimate the rate of false positives by disregarding sources of bias often found in biological sequences. An alternative approach is the use of reversed biological sequences as decoys, because these preserve the distribution of letters and the existence of tandem repeats, while disrupting the original sequence's functional properties. However, even randomly shuffled sequences contain palindromic elements that are on average longer than the longest common substrings shared between shuffled variants of the same sequence, leading to concerns that reversed sequences may retain signal of relationship that has been thought to be destroyed during reversal. Motivated by concerns that this may cause the empirical false positive rate of sequence annotation tools to be overestimated in benchmarks containing reversed sequences, we investigate the distribution of palindrome lengths in sequences and of alignment scores involving reversed sequences. We demonstrate that alignment between sequences and their reversals, even shuffled sequences with no notable repetitive or low-complexity regions, produces higher scores than alignment between truly unrelated sequences. We further demonstrate that this elevated score of alignment to reversals holds true even when the reversed sequences are subjected to moderate levels of mutation. While our analysis is primarily focused on the task of sequence annotation, we also demonstrate that the prevalence of internal palindromes leads to an overstatement of the rate of false labels in protein identification with mass spectrometry. As a result of these observations, we advise that reversed biological sequences be used as decoys only when care is taken to remove positive matches in the original (unreversed) sequences, or when overstatement of false labeling is not a concern.

sequence annotation | alignment | mass spectrometry | benchmark | reversal | false positive

Correspondence: ***georgeglidden@arizona.edu***

Introduction

The utility of tools that label biological sequences is dependent on the ability to reasonably predict the risk that a proposed label is due to some random or biased process, rather than true signal. An example is the case of sequence homology detection, in which an unlabeled sequence q is compared to a collection of known sequences, and similarity to one or more of these sequences serves as the basis of annotation of q. Homology detection tools typically represent the risk of false annotation by providing the supporting alignment's E-value along with its score S. For search of a particular query sequence against a given target database T, the E-value gives the number of alignments expected to meet or exceed the score S, if T consisted exclusively of unrelated (random) sequences. Theoretically-grounded calculation of the E-value has been established for gap-free alignments (1, 2), and empirical studies demonstrate that these E-value predictions are accurate for search against randomly generated sequences of nucleotides or amino acids (3, 4).

Even so, genomics practitioners are keenly aware that these E-values are often overestimates of the confidence with which we should label one sequence as being related to another, so that it is common for analysis pipelines to require seemingly-extreme E-value cutoffs for search matches (5). This is because the models underlying the scoring schemes used in alignment (6, 7) assume that the database sequences are generated under a simple random distribution (sample with replacement), whereas true biological sequences are not always random in this way, and can be readily distinguished from randomized noise.

One notable deviation comes in the form of locally repetitive sequence. So-called tandem repeats are typically the result of some form of replication error (8), and can present with a repeat period ranging from 1 to several hundred (9). They produce substantial risk of spurious matches when both query and target sequence contain (possibly degenerate) repeats with short period. One strategy for compensating for tandem repeats is to employ a tandem repeat *masking* tool (10–12) to hide regions of repetitive sequence away from homology detection. Such strategies are necessarily limited by the selection of some score threshold, and may thus either over-mask (hiding true homologs) or under-mask (allowing decayed repetitive regions to be identified as true homologs (13)).

Continuing concerns about real-world accuracy of E-values, and of the compensatory score adjustment/masking procedures, have lead researchers to evaluate tools with benchmarks that *empirically* measure false positive annotations in addition to true positive recall, rather than rely solely on theoretical E-values. A benchmark will generally contain a set

of *query* sequences and a distinct set of *target* sequences. In the target set, some sequences are related to a query (these are the *positives* that the tool will ideally annotate as such), while other sequences are not related to a query (these are *negatives* that the tool should not annotate). The benchmarked algorithm is used to compute alignments without knowledge of true labels, and a predicted classification is produced. Wherever the computed classification agrees with ground truth, the algorithm is said to have produced a *true positive* or *true negative*, and elsewhere it produces *false negatives* and *false positives*, respectively.

A variety of techniques have been developed to automate the production of positive and negative sequence pairs. In this paper, we are particularly interested in methods of creating negative sequence pairs, which we call *decoys*. One option for generating *decoy* sequences is to randomly sample real biological sequences from a large dataset. Importantly, there is a nontrivial chance of encountering genuine homology between a randomly-sampled pair of sequences, resulting in an overestimate of the rate of false positives and potentially biasing the algorithm to ignore true homologies. Alternative methods for producing *decoys* include (i) shuffling the letters of a real biological sequence, or (ii) generating a sequence by sampling letters under a random model. In both of these cases, failure to include repetition or other realistic features of biological sequences may lead to an underestimate of the rate of false positives.

Another mechanism for producing non-homologous biorealistic sequence is to use reversed biological sequences, so that decoys are generated by selecting true biological sequences and reversing the order of their characters. Reversal has several immediate benefits over stochastic methods. Reversal preserves the distribution of letters in a sequence, and also preserves the existence of short (possibly degenerate) tandemly repetitive regions and regions of low compositional complexity that are often found in biological sequences. Meanwhile, the reversed sequence almost certainly does not look or behave the same as the original. In the case of proteins, there is essentially no chance that a protein molecule from the reversed sequence will yield the same structure as the original unless that sequence is palindromic. In DNA, most codons are not palindromes, and other factors such as binding sites are likely to be disrupted by reversal.

Thus, a decoy set consisting of reversed sequences is expected to possess the low complexity and repetitiveness of real biological sequence, but with no true homology to an actual sequence. Such sequences demonstrate much higher rates of apparent "false" matching than do shuffled sequences, and it is tempting to view these as a better (more realistic) benchmark for false labeling, in the context of both nucleotide (11) and protein (14) annotation. We have employed reversed sequences in this way (15) for the same reasons; however, during post-publication analysis of our benchmark results, we realized that "false" matches of transposable element queries to a reversed (not complemented) genome were surprisingly likely to be to "self" – a match of a sequence query to the reversed genome was likely to be to a reversal of an instance its own family. In an earlier paper evaluating software for protein search, Eddy (16) suggested the source of this observation ("reversed sequences are surprisingly significantly more likely to show a significant match to the original sequence [...] because of a counterintuitive statistical effect of the frequency of approximate palindromes").

Here, we explore this counterintuitive phenomenon, and the extent to which it may cause the use of reversed decoys to overstate the risk of false discovery in sequence database search tools. We are not aware of any published analysis of this phenomenon. Through empirical analysis, we demonstrate that this palindromic phenomenon explains much of the increase in false discovery relative to shuffled sequence. We also note that reversed sequences are commonly used in developing the decoy sets that are used to tune mass spectrometry score thresholds to meet a target false discovery rate (17), and demonstrate that internal palindromes leads to an overstatement of the risk of false label rates in that domain.

As a result, we suggest that care be taken in using reversed sequences as a decoy sequence dataset when evaluating the occurrence of false discovery (18) or selecting score thresholds for reporting (13). If reversed sequences are to be used as decoys, they should be filtered to remove sequences that are matches to the query set prior to reversal. Alternatively it may be appropriate to use simulated sequences in which care is taken to produce realistic sequence (for DNA, see (19)). We hope that this manuscript will induce exploration into alternative mechanisms for improving the generation of biorealistic decoy sequences (particularly proteins) for approximating false similarity rates.

Intuition and Background for Palindrome Frequencies. We find it instructive to consider a simple calculation for the expected frequency of a length-n palindrome in a string A, and of a length-n common substring in two strings A and B, both over a uniformly distributed alphabet of size k. Consider some position i in A with value a_i. For reasonably large random strings, with modest value k, there will almost certainly be some position j in B with value b_j=a_i; this is a length-1 common substring. Ignoring edge effects, the chance that this is the center position of a length-n common substring (where n is odd) is $(1/k)^{(n-1)}$, because each neighboring character in A has a $1/k$ chance of being matched by the corresponding position in B.

Contrast this with the chance of observing a length-n palindrome. Consider a length-m string A and its reverse $\bar{A}$. Position i in A is the same letter as position (m-i+1) in $\bar{A}$. The chance that this is the center position of a length-n palindrome is $(1/k)^{(n-1)/2}$ – if we freeze the $(n-1)/2$ characters to the left of i, then a length-n palindrome is formed if the $(n-1)/2$ characters to the right are matched in reverse, each with $1/k$ probability. That is: in a randomly sampled sequence A, a length-n palindrome is less surprising than is a length-n substring shared by A and some other randomly sampled sequence B (or a shuffled variant of A). These anal-

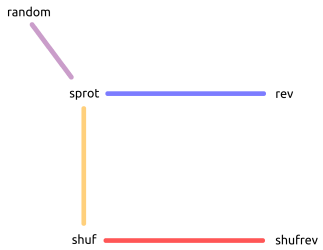

Fig. 1. The fundamental experiment design compares five related sequence sets, **random**, **sprot**, **rev**, **shuf**, and **shufrev**, represented as nodes, by means of pairwise alignment. Each colored edge represent the space of common subsequences between the the paired sequences of the sets.

yses involve odd-length matches, but transfer trivially to even lengths.

More generally, the expected length of the longest common sequence (LCS) of two strings of length-n is $2\log_{1/\lambda_2}(n)$ (20), and the expected longest palindrome length in a length-n string is $2\log_{1/\lambda_2}(n)+1$. (The probability amplitude λ_2 is determined from frequencies of each letter; for an alphabet of size s, on uniformly-distributed sequences, $\lambda_2 = 1/s$).

Finally, we consider approximate palindromes. An α-gapped palindrome is a kind of approximate palindrome defined as $w = uv\bar{u}$ where u and v can be any strings and $\bar{u}$ is the reversal of u. We say w is α-gapped whenever $\alpha \geq \frac{|u|+|v|}{|u|}$. In their 2017 paper, Duchon et al. (21) give an expectation for the length of α-gapped palindromes (L_n) as $\mathbb{E}[L_n] \sim (\alpha+1)\log_{1/\lambda_2}(n)$ or equivalently $\mathbb{E}[L_n] \sim (\beta+2)\log_{1/\lambda_2}(n)$ where $\beta = \alpha+1 \geq \frac{|v|}{|u|}$. By considering odd-length palindromes as gapped palindromes with $|v| = 1$, it follows that $\beta\log_{1/\lambda_2}(n) \gtrsim 1$, and it can be demonstrated that the expected length of odd maximal palindromes becomes $\mathbb{E}[L_n] \gtrsim 2\log_{1/\lambda_2}(n)+1$ [manuscript in prep]. To our knowledge, there is no further published analysis of approximate palindromes involving substitutions or a more general pattern of gap distributed throughout the palindrome.

Outline. In the following sections, we explore the connection between approximate palindrome length and the increased scores produced when aligning sequences against their reversal, relative to aligning against shuffled sequences. We begin by describing the sequence data used in our analysis, then examine the extent to which sequences achieve higher than random score when aligned to their reversal. We follow this with a study of the frequency of long palindromes and common substrings, and and investigation of the the roles of tandem repeats and mutations in the context of using reversed sequences as decoy sequences. Finally, we explore the impact of palindromes on assessing the false discovery rate in mass spectrometry analysis.

Methods

Preparation of Datasets.

Real protein sequences and their paired relatives. To gain insight into the influence of decoy type on the distribution of high-scoring matches between sequences, we developed related sets of paired protein sequences. For each pair, we first identified one sequence q from Swiss-Prot (22), then either (a) selected another Swiss-Prot sequence t, (b) shuffled q, or (c) reversed q.

Because random pairing of q and t can result in some pairings of true homologs, we aimed to reduce this risk by ensuring that paired sequences did not share folds as defined by the SCOP database (23, 24). Specifically, sequences from release 2022_04 of Swiss-Prot were length-filtered to identify 565,854 sequences with length between 50 and 2000; this set is called $\textbf{sprot}_{\textbf{all}}$. These were aligned to all SCOP2 superfamily representatives using blastp from the BLAST suite (version 2.12.0+) (25). Swiss-Prot sequences were retained for this dataset only if they yielded an alignment that covered at least 95% of the length of both the Swiss-Prot and SCOP sequence – this ensured that we could reasonably annotate all folds on retained proteins. The result is a set of 106,740 sequences, called (**sprot**). Length distributions are shown in Figure 2.

Sequences in (**sprot**) were assigned to clusters according to their SCOP superfamily. For each sequence q in **sprot**, its cluster v was identified, and a distinct cluster $w \neq v$ was chosen via sampling weighted by cluster size. A sequence t was then sampled from w and assigned to **random**. This second set (**random**) contains presumably non-homologous partners for the set **sprot**, and can be used to produce an alignment score distribution by aligning the ith sequence of **sprot** to the ith sequence of **random**.

To explore the distribution of scores when aligning a biological sequence to a randomly shuffled biological sequence, each sequence in **sprot** was mono-nucleotide shuffled to produce the set **shuf**; the ith sequence in **shuf** will share length and amino acid composition with the ith sequence in **sprot**, but no homology. Similarly, each sequence in **sprot** was reversed to produce the set **rev**. All shuffling and reversals were computed using the esl_shuffle tool from the Easel package as released in HMMER 3.3.2 (26).

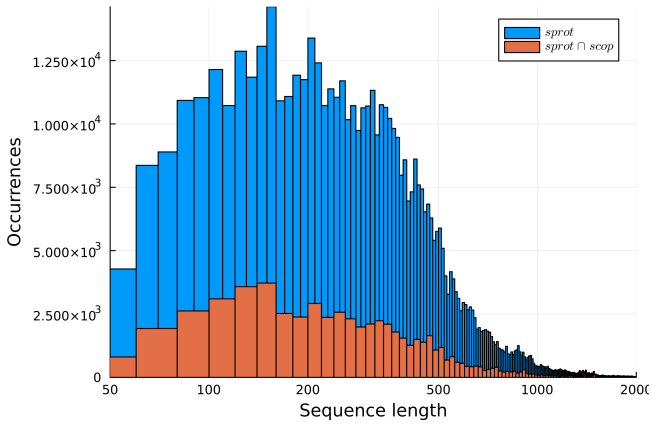

Fig. 2. Distribution of sequence lengths in Swiss-Prot.

It could be argued that the high rate of palindromes in biological sequences is due to repetition in those sequences. To demonstrate that the palindrome phenomenon is general, we reversed the sequences of **shuf** to produce **shufrev**; because the shuffled sequences do not retain signal of repetition found in true proteins, an alignment of a sequence in **shuf** to its mate in **shufrev** will highlight pure features of reversal alignment. Analogous sets of unrestricted relatives of **sprot$_{\mathbf{all}}$** are called **rev$_{\mathbf{all}}$**, **shuf$_{\mathbf{all}}$**, and **shufrev$_{\mathbf{all}}$** respectively.

Reversals in mutated sequences. In the reversal analyses described above, each **sprot** sequence is compared to a reversal of itself. In a carefully constructed benchmark, query sequences themselves may not be present in the reverse decoy set, calling the applicability of this analysis into question. But the benchmark may contain their homologs, which will show some level of sequence similarity to the queries. To understand how the level of divergence impacts distribution of scores induced by approximate palindromes, we generated six mutated variants of **shuf$_{\mathbf{all}}$**, and reversed each to form a new diverged reversed benchmark. For each variant, a conditional distribution was used as the basis of mutation, so that $s = s_1 \cdots s_n$ in **shuf$_{\mathbf{all}}$** was transformed to a related sequence $t = t_1 \cdots t_n$, in which t_i is an amino acid sampled from the distribution $p(x|s_i)$. For the first transformation, substitution probabilities were derived (27) from the BLOSUM90 substitution matrix. This matrix has a mean self-emission frequency ($p(a|a)$) of 45.6%. Additional transformations were produced based on modified distributions in which the likelihood of mutation was reduced for each amino acid a, adjusting the likelihoods of mutation to the other 19 amino acids $p(b|a)$ proportionately. These distributions have average self-emission frequencies ($p(a|a)$) of 50%, 60%, 70%, 80% and 90% respectively. These enable consideration of the likelihood of finding matches between a query and reversals of homologs that show varying levels of percent identity to the query. These augmented distributions preserve the relative frequency of mutations derived from BLOSUM90 ($p(b|a)$ for $b \neq a$) while increasing the expected percent identity between the original sequence s and its relative t. We refer to these distributions as **shufrev$_{\mathbf{all}}$-46pct**, **shufrev$_{\mathbf{all}}$-50pct**, **shufrev$_{\mathbf{all}}$-60pct**, **shufrev$_{\mathbf{all}}$-70pct**, **shufrev$_{\mathbf{all}}$-80pct**, and **shufrev$_{\mathbf{all}}$-90pct**, respectively.

Repeat Masking. To further investigate the role of repetition in false matching for biological sequences, we hard-masked repetitive sequences in **sprot$_{\mathbf{all}}$** using *tantan* version 40 (28) with default parameters, producing **masked sprot$_{\mathbf{all}}$**.

Simulated fragmentation with trypsin. In mass spectrometry-based proteomics, proteins are broken into fragments, often employing trypsin digestion. The identity of the resulting fragments is determined based in part on their mass. Typically, the fragments evaluated by mass spectrometry have length 5 to 100 amino acids because the signal induced by short sequences (< 5 residues) is indistinguishable from noise, while long sequences (> 100 residues) may exceed the the physical limits of the machine. To simulate this process, we applied a simplified approximation of trypsin digestion as follows. Consider a protein $P = (p_1, p_2, ..., p_n)$ and its set $S = (s_1, s_2, ..., s_k)$ of potential trypsin digestion positions at which there is an arginine or lysine, ignoring any position that is not followed by a proline. Taking a pair of positions (s_i, s_j) from S, P may be split into three fragments, $(p_1..p_{s_i}), (p_{s_i+1}..p_{s_j}), (p_{s_j+1}..p_n)$. For all protein sequences in **sprot$_{\mathbf{all}}$**, we enumerated all pairs (i, j) of potential digestion sites and produced all resulting three-way splits, then filtered to remove fragments with length <5 or >100. The remaining fragments were accumulated into the set **sprot$_{\mathbf{tr}}$**. Each member of **sprot$_{\mathbf{tr}}$** was shuffled to produce a shuffled control denoted **shuf$_{\mathbf{tr}}$**. This approach does not capture the true complexity of trypsin digestion patterns (29), but is sufficient for our purposes of analyzing palindromes in relatively short fragments.

Annotation of Chromosome 22. The preceding datasets only comprise protein and peptide sequences. To demonstrate that the elevated presence of approximate palindromes in sequence benchmarks is not merely a feature of amino acid sequences, we consider DNA sequences from the human genome.

The sequence of human chromosome 22 in hg38.p14 along with the NCBI RefSeq annotation was downloaded from UCSC's Genome Browser (30). The command-line tool GffRead (31) was used to extract and concatenate exon regions into sequences corresponding to the mRNA transcripts of chromosome 22, denoted **mRNA**. The intron and intergenic regions of chromosome 22 were also sampled for sets of sequences with similar lengths as **mRNA**, denoted **intron** and **intergenic**.

As above, shuffling and masking were performed, and the resulting variant datasets are denoted **shuf$_{\mathbf{mRNA}}$**, **shuf$_{\mathbf{intron}}$**, **shuf$_{\mathbf{intergenic}}$**, and **masked mRNA**, **masked intron**, and **masked intergenic**.

Software and algorithms.

Exploration of score and count distributions was achieved using established tools and algorithms. For the large Swiss-Prot-based dataset, sequence database search scores were computed using *blastp* (25). For our purposes, pairwise alignment was performed by creating a blast database for each singular target, and this target was "searched" with its paired query sequence. The parameter for (minimum) word length was set to 4. The default substitution matrix (BLOSUM62) and gap penalties (gap open $= -11$, gap extend $= -1$) were left unchanged. One *blastp* alignment may contain several results, each with its own score – for our experiments, only the highest-scoring match was retained and reported. We report score, rather than the associated E-value, because E-value is dependent on the size of the query and target datasets, while score is stable for an aligned pair.

One problem with the BLAST-based analysis is that for many comparisons, the similarity between sequences is sufficiently poor that not even a low-scoring *blastp* result is produced. To

gain insight to the distribution of lower-scoring results, we used the BioJulia implementation of Smith-Waterman (SW) algorithm BioAlignments.jl (32) for finding the optimal local alignment between two sequences. The algorithm is parameterized by a substitution matrix (whose values at row i and column j determine the cost of substituting letter i with letter j), as well as gap-open and gap-extend penalties, which determine the cost of skipping one or more characters in either sequence. When gap penalties and off-diagonal values in a substitution matrix are very large negative numbers (e.g., $-1e12$), Smith-Waterman will never produce gaps or mismatches and effectively performs exact alignment; we use this mode for computing longest common substrings (LCS) between sequences. In all other cases the gap penalties and substitution matrix are set to BLAST's defaults.

Using either BLAST or Smith-Waterman (with typical parameters) for reversed alignment readily produces approximate palindromes, but rarely identifies true palindromes. To measure the frequency of exact palindromes in Swiss-Prot, we implemented Manacher's algorithm (33) in Julia.

Code used to produce all results and figures may be found in Supplementary Material at `https://github.com/TravisWheelerLab/WASITAMATchISAW`.

Results

Alignment scores for Swiss-Prot variations. To understand the influence of reversals on sequence alignment score distributions, we aligned sequences taken from Swiss-Prot (length between 50 and 2000) against a variety of decoy datasets. Each pairing involves either 106,740 (**sprot** and relatives) or 565,855 (**sprot$_{\text{all}}$** and relatives) pairwise alignments, with the ith sequence of one set aligned to the ith sequence of the other set (using *blastp*). The first pairing (**sprot↔random**) replicates the case in which a real protein sequence is aligned to some randomly chosen (presumed non-homologous) protein. The second pairing (**sprot↔shuf**) corresponds to a simple decoy strategy: the true protein is aligned to a sequence that is a shuffled copy of the original – it shares length and amino acid composition, but is not a true protein. As seen in Figure 3A, the distribution of scores in these two pairings are similar.

The next two pairings involve reversed sequences. The **sprot↔rev** pairing aligns each of the 106,740 true Swiss-Prot proteins to the sequence's reversal. Figure 3 shows that the distribution of these scores is strongly right-shifted relative to alignments involving **random** or **shuf**. According to common wisdom this is expected, as the argument for using reversed sequences as decoys is that real biological sequences contain patterns of repetition and regions of low complexity, and these patterns are responsible for a higher observed rate of matches against reversed decoys. This result is contrasted with the pairing **shuf↔shufrev**. In this pairing, each query sequence is a shuffled sequence from Swiss-Prot, and the target for each shuffled query is its reversal. The shuffled sequences retain only random levels of repeats and local low complexity, so by conventional wisdom are expected to show a score distribution that is similar to that of **sprot↔shuf**. However, the score distribution of **shuf↔shufrev** is nearly identical to that of the biological sequence and reversal (Figure 3A (for **sprot**) and 3B (for **sprot$_{\text{all}}$**). This indicates that the observed right-shift of **sprot↔rev** scores is primarily due to the reversal itself, and not the biological or repetitive nature of the sequences.

Figure 3C reproduces the reversal analyses from panel B, but for only sequences at least 500 in length (107205 total sequences). In both 3B and 3C, the distribution of **sprot$_{\text{all}}$↔rev$_{\text{all}}$** scores is slightly right-shifted relative to **shuf$_{\text{all}}$↔shufrev$_{\text{all}}$**, and the feature is more prominent in B. This shift likely represents the modest influence of repetitive and low-complexity sequence on scores of alignment to reversed sequences.

Figure 4 presents scores for these various pairings as a function of length. Specifically, every pair was binned according to sequence length, then the mean *blastp* score was computed for each bin. Sequences aligned to other **sprot** sequences (**sprot↔random**) may differ in length, so were binned by the geometric mean length of the query and target sequences; they show a score distribution that is essentially identical to that of sequences aligned to their shuffled match (**sprot↔shuf**). When **sprot** sequences are aligned to their reversals (**sprot↔rev**), they produce consistently higher scores than they do when aligned to their shuffled partners (**sprot↔shuf**), but nearly all of the difference is explained by the simple fact of reversal, as evidenced by the gap between **sprot↔shuf** and **shuf↔shufrev** and the overlap between **sprot↔rev** and **shuf↔shufrev** and .

Role of Tandem Repeats in BLAST Score Gap. The argument for using reversed sequences as decoys in benchmarking is that they present a challenge because they can highlight the risk of false matches due to (possibly degenerate) repetitive sequence. Because Swiss-Prot sequences aligned to their reversals show a modest increase in BLAST score relative to shuffled sequences to their reversals, we sought to explore the role of tandem repeats in this shift. We created the dataset **masked sprot$_{\text{all}}$** by hard-masking repetitive regions. Figure 4B presents these analyses, focused on **shuf$_{\text{all}}$↔shufrev$_{\text{all}}$** and **sprot$_{\text{all}}$↔rev$_{\text{all}}$** distributions, while Figure 4C replaces **sprot$_{\text{all}}$↔rev$_{\text{all}}$** with another distribution showing the result of first masking with *tantan* (28) (default settings), then searching against its reversal (**masked sprot$_{\text{all}}$↔rev$_{\text{all}}$**). These results demonstrate that the small gap between **shuf$_{\text{all}}$↔shufrev$_{\text{all}}$** and **sprot$_{\text{all}}$↔rev$_{\text{all}}$** is almost entirely due to repetitive sequence that can be effectively masked by a tool such as *tantan*.

The Role of Mutations (Approximate Reversals). The above methods have only examined alignments between a sequence and its exact reversal. In practice, a database of reversed sequences may contain the reversal of a homolog to the query sequence, in which case we expect the effect of reversal on alignment score to be diminished. To simulate

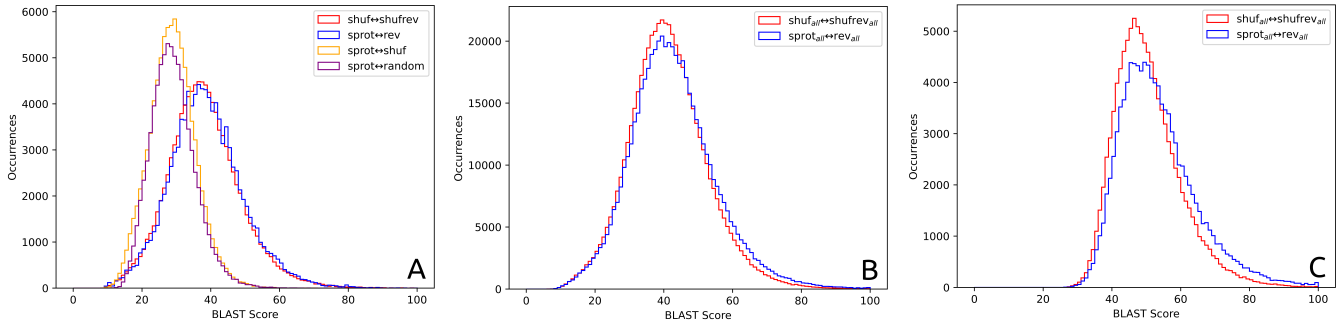

Fig. 3. Score distributions from BLAST alignments. (A) All four pairings **sprot↔random**, **sprot↔shuf**, **sprot↔rev**, and **shuf↔shufrev** are restricted to sequences with 95% identity to SCOP2 superfamily representatives. (B) The unrestricted scores of **sprot$_{all}$↔rev$_{all}$** and **shuf$_{all}$↔shufrev$_{all}$** are shown. (C) **sprot$_{all}$↔rev$_{all}$**, and **shuf$_{all}$↔shufrev$_{all}$** are restricted to sequences with length greater than or equal to 500 residues.

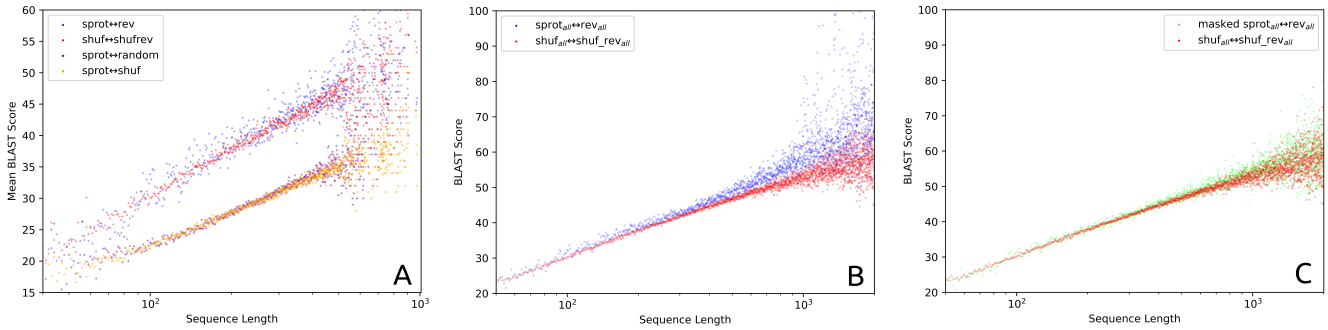

Fig. 4. BLAST alignment scores as a function of sequence length, for the set of 106,740 sequences with full-length matches to SCOP sequences. Sequences were binned according to sequence length, and mean blastp score was computed for each bin. (A) The set **sprot↔random** includes pairs that may not have equal length, so scores are plotted by the geometric mean length of query and target sequence. For all remaining sets both sequences in each pair share the same length: **sprot↔shuf**, **sprot↔rev**, and **shuf↔shufrev** are plotted. (B) This plot shows scores as a function of sequence length for all 565,854 Swiss-Prot sequences with length under 2000: **sprot$_{all}$↔rev$_{all}$**, and **shuf$_{all}$↔shufrev$_{all}$**. (C) The values from B are reproduced, and supplemented with the results of masking with tantan prior to search (**masked sprot$_{all}$↔rev$_{all}$**); the resulting distribtuon shows that most of the modest difference between **sprot$_{all}$↔rev$_{all}$** and **shuf$_{all}$↔shufrev$_{all}$** is resolved by application of tandem repeat masking.

this situation, we produced a collection of reversal variants by sampling amino acid substitution mutations from conditional distributions with self-emission probabilities ranging from 46% to 90%. These transformation are of **shufrev$_{all}$**, named **shufrev$_{all}$-46pct**, **shufrev$_{all}$-50pct**, **shufrev$_{all}$-60pct**, **shufrev$_{all}$-70pct**, **shufrev$_{all}$-80pct**, and **shufrev$_{all}$-90pct**. Pairwise alignments are performed for each transform to **shuf$_{all}$** using the Smith-Waterman (SW) algorithm implemented in *BioAlignments.jl* (32).

The SW alignments between **shuf$_{all}$** and **shufrev$_{all}$-Xpct** for X ∈ {46, 50, 60, 70, 80, 90} were computed using the BLOSUM62 scoring matrix and gap penalties set to −11 for opening and −1 for extension, to mimic the default configuration of *blastp*. Figure 5 displays the spread of these scores, as well as those of **sprot$_{all}$↔shuf$_{all}$** and **shuf$_{all}$↔shufrev$_{all}$**. The self-matching effect of reversal is lost completely in **shufrev$_{all}$-46pct** and almost undetectable in**shufrev$_{all}$-50pct**, which present score distributions similar to **sprot$_{all}$↔shuf$_{all}$**. But when simulated homologs are created with higher average percent identity between query and partner, and the partner is then reversed (yielding **shufrev$_{all}$-Xpct** for X ∈ {60,70,80,90}), the associated pairs exhibit higher-than-random affinity between sequences and their (approximate) reversal. Thus, there is a risk of overstating false match rates even when the reversed decoy set contains homologs sharing 60% or greater similarity to query sequences.

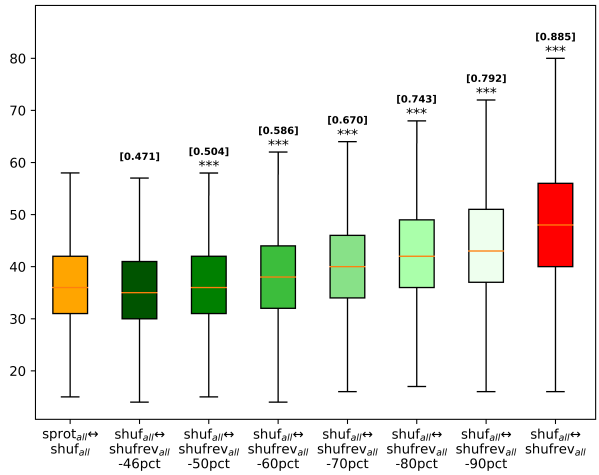

Fig. 5. Box-and-whisker plot of SW alignment score distributions for alignment pairs **sprot$_{all}$↔shuf$_{all}$**, **shuf$_{all}$↔shufrev$_{all}$**, and **shufrev$_{all}$-Xpct** for X ∈ {46, 50, 60, 70, 80, 90}. Above each box plot, the symbol *** indicates distributions that are significantly higher than **sprot$_{all}$↔shuf$_{all}$** based on the Wilcoxon rank test (P = 0, due to computational underflow). The values [Y] indicates the fraction of shuffled instances for which the associated box plot produces a higher score than **sprot$_{all}$↔shuf$_{all}$**.

Exact palindromes in Swiss-Prot. BLAST is sensitive to inexact alignments and makes biologically-motivated substitutions where one peptide may naturally be replaced by another. While this is appropriate for detecting homology in noisy real-world data, the output on alignment to reversal rarely

produces a perfect palindrome. To better understand the reversal-invariant structures present in Swiss-Prot proteins, we computed all maximal-length palindromes in each of the 565,855 sequences in $\mathbf{sprot_{all}}$ as well as $\mathbf{shuf_{all}}$ for a control, using our implementation of Manacher's algorithm (33). An additional control was observed by computing the longest common substring (LCS) between sequence pairs in $\mathbf{sprot_{all}}$ and $\mathbf{shuf_{all}}$.

The palindromic substrings in $\mathbf{sprot_{all}}$, $\mathbf{shuf_{all}}$, and the LCS between them, were binned by the length of the sequence containing the substring; averages computed per each length are presented in Figure 6. The length distribution of maximal palindromes is closely related to that of longest common substrings; a result of Duchon et al. (21) implies that the expected length of (ungapped) maximal palindromes is asymptotically equivalent to the expected length of longest common substrings as derived by Arratia et al. (20). In order to contextualize our analyses within this theory, we plotted the expected lengths of (i) LCS between two random sequences ($2\log_{1/\lambda_2}(n)$), and (ii) palindrome within a sequence ($2\log_{1/\lambda_2}(n)+1$).

Figure 6 shows that $\mathbf{sprot_{all}}$ contains, on average, longer palindromic substrings than $\mathbf{shuf_{all}}$. While palindromes found in the shuffled dataset are densely concentrated around their expected value, the average palindrome length in $\mathbf{sprot_{all}}$ diverges from the expectation, particularly as sequence length increases. This asymptotic behavior only yields a large divergence from expected score in the small fraction of sequences in $\mathbf{sprot_{all}}$ that are longer than $\sim$500 amino acids.

To explore the role of repetitive sequence in this shift, we computed maximal palindromes for each *masked* sequence. The resulting distribution of palindrome lengths (Figure 6B) mirrors that of the shuffled sequences in Figure 4C, clustering densely around the expected value. Based on manual review, the divergence of biological palindromes from their stochastic counterparts appears to be almost entirely explained entirely by long tandem repeats. In particular, these palindromes are composed of smaller, palindrome-supporting subunits like 'PNAN' or 'SD', leading respectively to palindromes like PNANPNANPNANP and SDSDSDSDS.

Palindromes in the context of mass spectrometry. When using mass spectrometry to identify proteins in a sample, a library of reversed sequences is usually used to set thresholds based on risk of false labeling. Here, we consider the potential that including reversals in mass spectrometry benchmarks may lead to overestimated false match rates and a resulting decrease in search sensitivity.

With all sequences in $\mathbf{sprot_{all}}$, we produced protein fragments similar to those used in mass spectrometry, splitting each sequence at each instance of a lysine or arginine, except where followed by a proline (see Methods). After splitting, substrings with length between 5 and 100 were accumulated into sets named $\mathbf{sprot_{tr}}$. Fragments in $\mathbf{sprot_{tr}}$ were shuffled to produce a related set $\mathbf{shuf_{tr}}$.

Figure 7 displays the average lengths of longest palindromic subsequence (LPS) in each set of strings, binned by the length of the string. To place these in context, the longest common substring (LCS) was computed between paired sequences in $\mathbf{sprot_{tr}}$ and $\mathbf{shuf_{tr}}$.

Trends established by the preceding results are still apparent in Figure 7. LPS are longer than LCS, and because tryptic peptides are so short, the constant component separating LPS and LCS lengths is large relative to their shared logarithmic component. However, the relationship between LPS and LCS appears to invert for very short sequences. The data indicate that $LCS(\mathbf{sprot_{tr}}, \mathbf{shuf_{tr}}) > LPS(\mathbf{sprot_{tr}})$ for sequences shorter than 10 amino acids, and $LPS(\mathbf{sprot_{tr}}) \sim LPS(\mathbf{shuf_{tr}})$ for sequences shorter than approximately 20

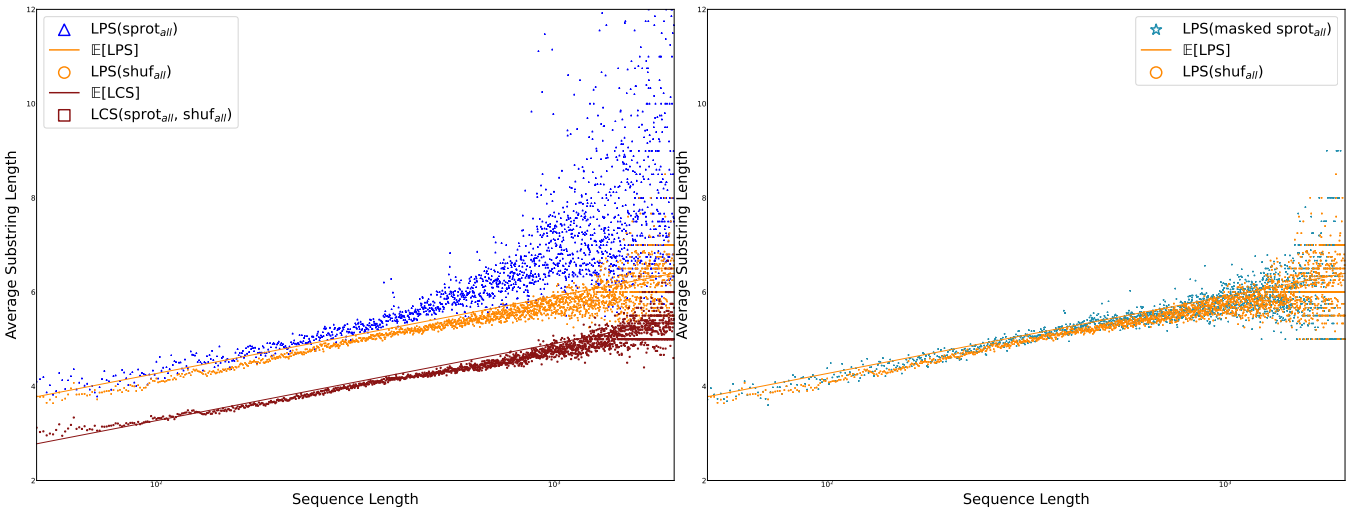

Fig. 6. Lengths of maximal palindromes occurring in biological sequences **sprot** and random sequences **shuf** with the expected lengths derived from (21). Biological palindromes appear to be asymptotically longer than random palindromes. On the left, the lengths of longest common substrings (LCS) of pairs between **sprot** and **shuf** are plotted with the expected length from (20). On the right, maximal palindromes remaining after masking **sprot** with *tantan* are plotted; with repetitive regions removed, the lengths of biological palindromes become closely clustered around the random expectation.

peptides. Composition bias in short tryptic peptides may contribute to quasi-invariance under shuffling.

Palindromes in DNA. The analyses presented above have been focused on protein sequences, but the surprising prevalence of palindromes is a general feature of strings, regardless of alphabet. As with the protein results shown in this manuscript, a random DNA sequence (i.e. one with no repetitive or low-complexity regions) will on average have a longer exact match to its reversal than it will to a shuffled version of itself; it will also typically produce a higher score when aligned (under a reasonable scoring scheme) to its reversal than when aligned to a shuffled version of itself.

To explore this, we analyzed DNA occurring in human chromosome 22. Chromosome 22 was downloaded and annotated using the UCSC Genome Browser. Gene sequences were extracted from the chromosome using GffRead; introns and intergenic regions were also sampled from chromosome 22 to match the length distribution of genes. These datasets are respectively denoted **mRNA**, **intron**, and **intergenic**. Each set was filtered to contain sequences less than 50000 bp. As above, variant sets were created using shuffling and masking, respectively denoted $\mathbf{shuf_{mRNA}}$, $\mathbf{shuf_{intron}}$, $\mathbf{shuf_{intergenic}}$, and **masked mRNA**, **masked intron**, and **masked intergenic**. For more detail, see the preceding section in Methods.

The LPS of each DNA sequence was computed using Manacher's algorithm. LCS were constructed between each sequence and a shuffled counterpart using a substring automata. Fig 8 plots the lengths of these common substrings, binned by sequence length.

The analysis of DNA reproduces the results observed in peptides and proteins. On average, the longest palindromic substring of a sequence is longer than the longest common substring between that sequence and a random other. When biologially-induced repetitions are present, the difference between LPS and LCS grows superlogarithmically. In the absence of repetition (either by shuffling or tantan masks), this difference appears constant.

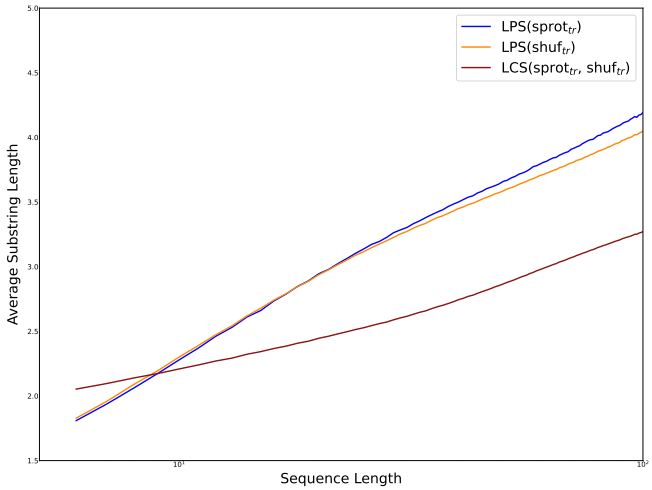

Fig. 7. The average length of LPS in tryptic peptides between lengths 5 and 100, generated from an in-silico trypsin digest of $\mathbf{sprot_{all}}$, are plotted against those of tryptic substrings occurring $\mathbf{shuf_{all}}$. When tantan-masked regions are removed as in **masked** $\mathbf{sprot_{all}}$, the difference in distributions of LPS length collapses.

Summary of Results. In summary, alignment to reversal produces a substantial increase in score compared to alignment to unrelated sequences, even when there is no inherent repetitiveness or low complexity in the initial sequence. This effect is diminished when the reversed sequence contains mutations relative to the query, but remains significant for related sequences exceeding 60% identity to the non-reversed sequence. As a result, when reversed biological sequences are used as benchmarks they may overstate the risk of false matches.

Discussion

Up to this point, we have discussed the prevalence of palindromes in terms of their impact on bioinformatic analysis (specifically, overstatement of false positive rates in sequence search). It is also interesting to consider how this existence of palindromes in sequences may play a role in biological processes. The role of palindromes in proteins is currently unknown and research on their function is ongoing (34–37). Sheari et al. (35) argue that the prevalence of palindromic regions in proteins is a side-effect of repetitive and low-complexity regions. Sridhar et al. (36) found that 60% of palindromes in proteins are "not associated with any regular secondary structure", consistent with the notion that palindromes in proteins are not dependent on repetitiveness such as that observed in alpha helical structures. Furthermore, 80% of palindromes in proteins do not interact with active sites, catalytic residues, ligands, or metal ions. In a more direct study, Aikaterini et al. (37) synthesized a 4-alpha-helical bundle and its reversal, and demonstrated that while the reversed sequence does fold into a helix – likely due to the aforementioned effects of repetition in helical sequences – the resulting structure was not a true alpha helix. In general, these results are consistent with our results above, in that the frequency of palindromes in true proteins appears to be primarily a function of the frequency of palindromes in random sequences, and repetitive aspects of some sequences may lead to a moderate increase in the observed rate of palindromes.

A better benchmark?. Can we compensate for the overabundance of palindromic regions in biosequence benchmarks? The supplemental materials include a script for generating improved reversed decoy benchmarks, given a query set Q and a target set T. The target set T is restricted to T', containing sequences with less than 50% percent identity to any sequence in Q. The reversal of T', denoted R, retains the original biological signal present in T and omits reversal-induced false matches. As above, it may be necessary to mask repetitive regions before performing the initial alignment.

ACKNOWLEDGEMENTS

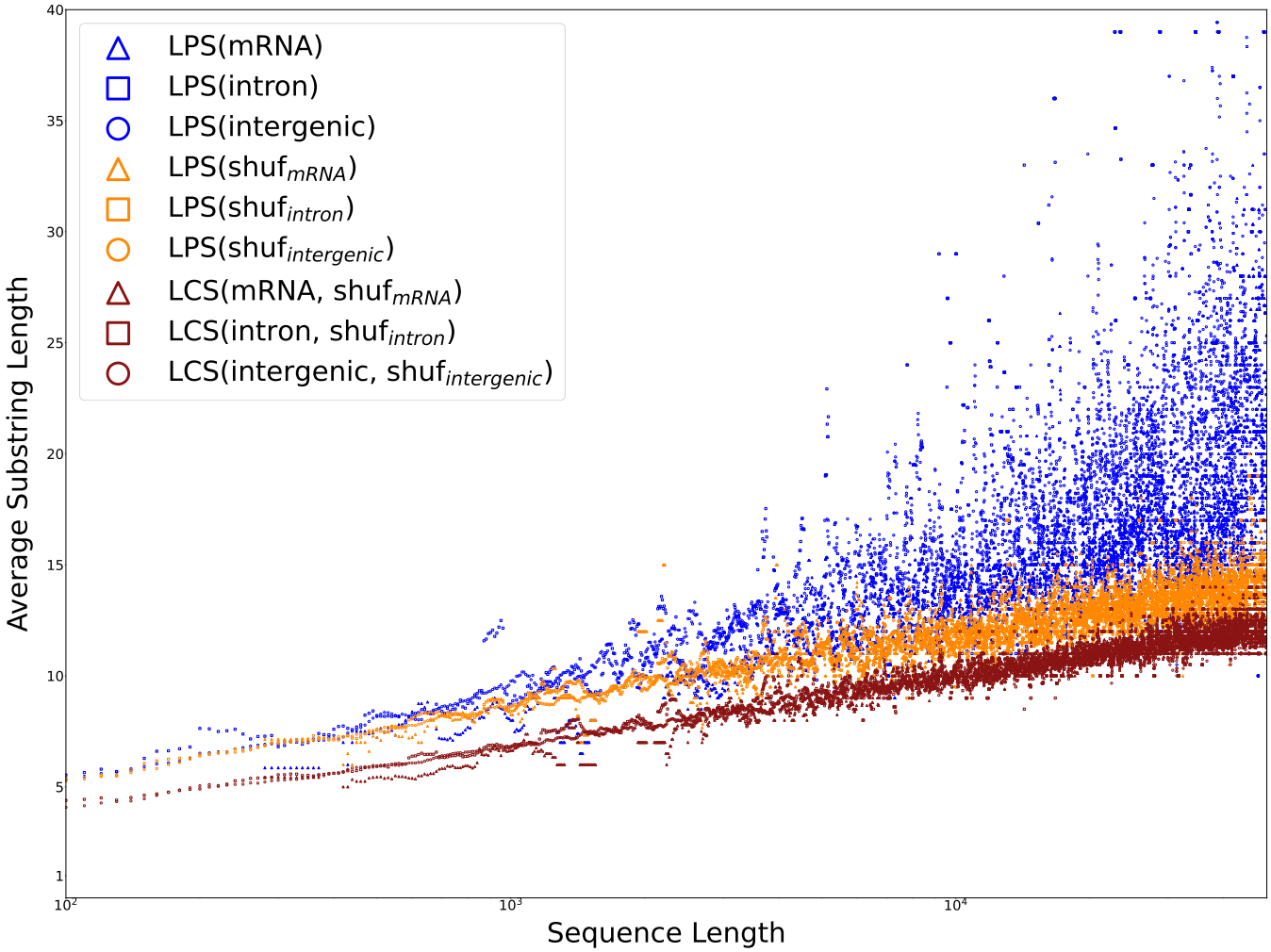

Fig. 8. The average lengths of palindromes occurring in human chromosome 22 are plotted by sequence length against shuffled controls.

We wish to thank Dave Rich for careful consideration of benchmarking issues during his thesis work, and for discussions that motivated early work on this analysis. We also thank Robert Hubley and Daniel Olson for illuminating conversations regarding both palindrome theory and application to annotation benchmarks. We also thank Martin Frith for providing valuable feedback and insights in response to our first preprint draft. Analyses were made possible thanks to high performance computing (HPC) resources supported by the University of Arizona TRIF, UITS, and Research, Innovation, and Impact (RII) and maintained by the UArizona Research Technologies department. GGH and TJW were supported by NSF DBI 1933305.

Bibliography

1. Samuel Karlin and Stephen F Altschul. 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, 1990.
2. Stephen F Altschul and Warren Gish. Local alignment statistics. In *Methods in enzymology*, volume 266, pages 460–480. Elsevier, 1996.
3. Richard Mott. Maximum-likelihood estimation of the statistical distribution of smith-waterman local sequence similarity scores. *Bulletin of Mathematical Biology*, 54(1):59–75, 1992.
4. Sean R Eddy. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS computational biology*, 4(5):e1000069, 2008.
5. William R. Pearson. An Introduction to Sequence Similarity ("Homology") Searching. *Current Protocols in Bioinformatics*, 6 2013.
6. Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.
7. Kimmen Sjölander, Kevin Karplus, Michael Brown, Richard Hughey, Anders Krogh, I Saira Mian, and David Haussler. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Bioinformatics*, 12(4):327–345, 1996.
8. Danilo Pumpernik, Borut Oblak, and Branko Borštnik. Replication slippage versus point mutation rates in short tandem repeats of the human genome. *Molecular Genetics and Genomics*, 279:53–61, 2008.
9. Jitendra Thakur, Jenika Packiaraj, and Steven Henikoff. Sequence, chromatin and evolution of satellite dna. *International Journal of Molecular Sciences*, 22(9):4309, 2021.
10. Gary Benson. Tandem repeats finder: a program to analyze dna sequences. *Nucleic acids research*, 27(2):573–580, 1999.
11. Martin C. Frith. A new repeat-masking method enables specific detection of homologous sequences. *Nucleic Acids Research*, 39(4):e23–e23, 11 2010. ISSN 0305-1048. doi: 10.1093/nar/gkq1212.
12. Daniel Olson and Travis Wheeler. Ultra: a model based tool to detect tandem repeats. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 37–46, 2018.
13. Jessica Storer, Robert Hubley, Jeb Rosen, Travis J Wheeler, and Arian F Smit. The dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*, 12(1):1–14, 2021.
14. Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.
15. Travis J Wheeler, Jody Clements, Sean R Eddy, Robert Hubley, Thomas A Jones, Jerzy Jurka, Arian FA Smit, and Robert D Finn. Dfam: a database of repetitive dna based on profile hidden markov models. *Nucleic acids research*, 41(D1):D70–D82, 2012.
16. Sean R Eddy. Accelerated profile hmm searches. *PLoS computational biology*, 7(10): e1002195, 2011.
17. James C Wright and Jyoti S Choudhary. Decoypyrat: fast non-redundant hybrid decoy sequence generation for large scale proteomics. *Journal of proteomics & bioinformatics*, 9 (6):176, 2016.
18. Maria Hauser, Martin Steinegger, and Johannes Söding. Mmseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics*, 32(9): 1323–1330, 2016.
19. Juan Caballero, Arian FA Smit, Leroy Hood, and Gustavo Glusman. Realistic artificial dna sequences as negative controls for computational genomics. *Nucleic Acids Research*, 42 (12):e99–e99, 2014.
20. Richard Arratia and Michael S Waterman. An erdös-rényi law with shifts. *Advances in Mathematics*, 55(1):13–23, 1985. ISSN 0001-8708. doi: https://doi.org/10.1016/0001-8708(85)90003-9.
21. Philippe Duchon, Cyril Nicaud, and Carine Pivoteau. Gapped pattern statistics. In *Annual Symposium on Combinatorial Pattern Matching*, 2017.
22. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1): D523–D531, 2023.

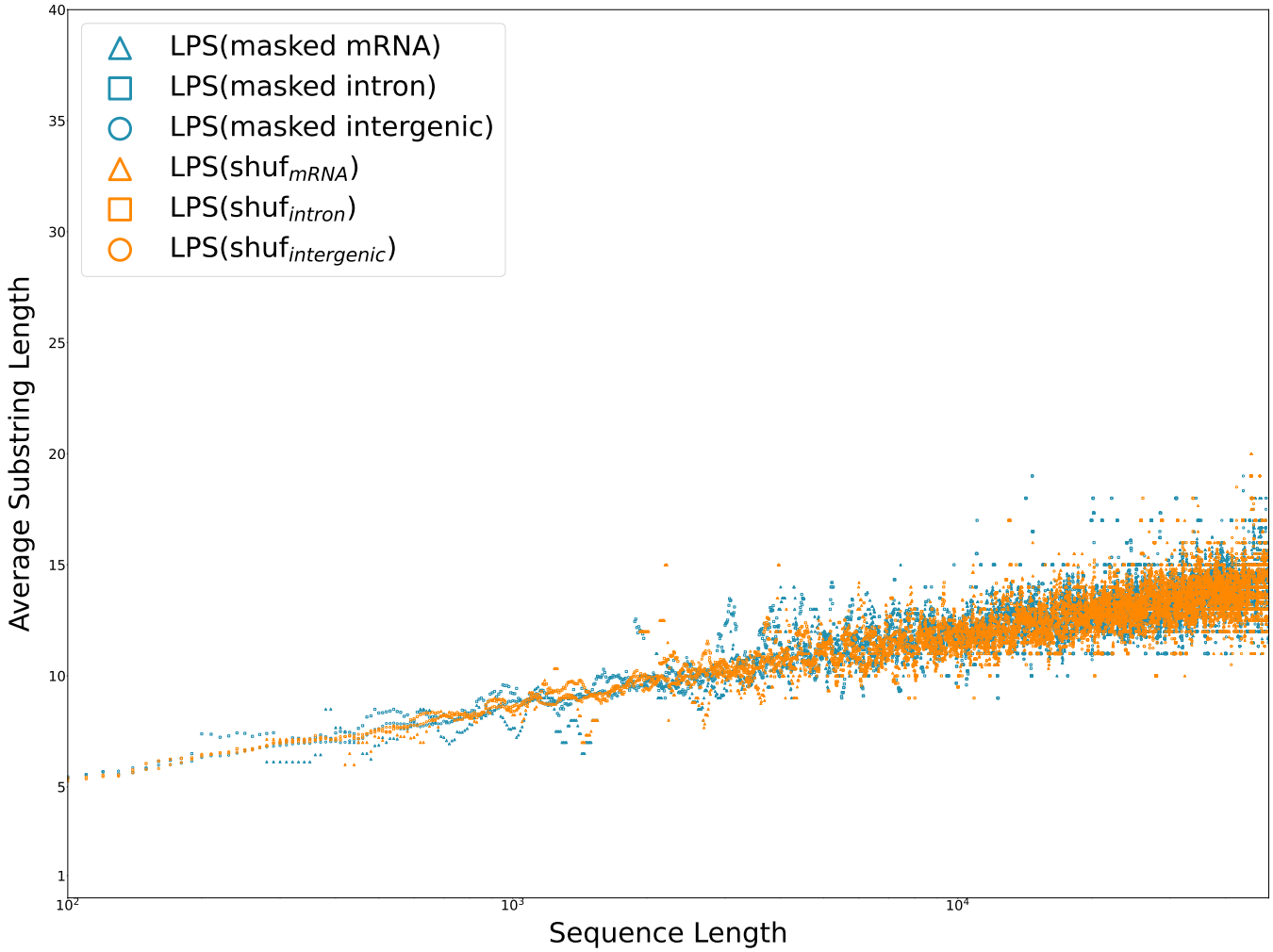

Fig. 9. Repetitive regions in human chromosome 22 are masked via tantan; the remaining palindromes in non-repetitive DNA, as well as those in corresponding shuffled sequences, are plotted. Both data are distributed closely around their expectation. As above, when repetitive subsequences are suppressed, the biological contribution to reverse affinity appears negligible.

23. Antonina Andreeva, Dave Howorth, Cyrus Chothia, Eugene Kulesha, and Alexey G Murzin. Scop2 prototype: a new approach to protein structure mining. *Nucleic acids research*, 42 (D1):D310–D314, 2014.
24. Antonina Andreeva, Eugene Kulesha, Julian Gough, and Alexey G Murzin. The scop database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic acids research*, 48(D1):D376–D382, 2020.
25. W James Kent. Blat—the blast-like alignment tool. *Genome research*, 12(4):656–664, 2002.
26. Robert D Finn, Jody Clements, and Sean R Eddy. Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, 39(suppl_2):W29–W37, 2011.
27. Yi-Kuo Yu and Stephen F. Altschul. The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. *Bioinformatics*, 21 7:902–11, 2005.
28. Martin C Frith. A new repeat-masking method enables specific detection of homologous sequences. *Nucleic acids research*, 39(4):e23–e23, 2011.
29. Jesse Rodriguez, Nitin Gupta, Richard D Smith, and Pavel A Pevzner. Does trypsin cut before proline? *Journal of proteome research*, 7(01):300–305, 2008.
30. W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.
31. Geo Pertea and Mihaela Pertea. Gff utilities: Gffread and gffcompare. *F1000Research*, 9, 2020.
32. Thomas A. Christensen II, Ciarán O'Mara, Alexey Stukalov, Kenta Sato, Jakob Nybo Nissen, Sabrina Jaye Ward, Kevin Bonham, Daniel C. Jones, Tim Holy, Kapara newbie, Morten Piibeleht, and tanhevg. Biojulia/bioalignments.jl: v3.1.0, January 2023.
33. Dan Gusfield. Algorithms on strings, trees, and sequences. 1997.
34. Malgorzata Giel-Pietraszuk, Marcin Hoffmann, Sylwia Dolecka, Jacek Rychlewski, and Jan Barciszewski. Palindromes in proteins. *Journal of protein chemistry*, 22:109–113, 2003.
35. Armita Sheari, Mehdi Kargar, Ali Katanforoush, Shahriar Arab, Mehdi Sadeghi, Hamid Pezeshk, Changiz Eslahchi, and Sayed-Amir Marashi. A tale of two symmetrical tails: Structural and functional characteristics of palindromes in proteins. *BMC bioinformatics*, 9:1–7, 2008.
36. Settu Sridhar, Mallapragada Nagamruta, and Kunchur Guruprasad. Analyses of the sequence and structural properties corresponding to pentapeptide and large palindromes in proteins. *Plos one*, 10(10):e0139568, 2015.
37. Aikaterini Kefala, Maria Amprazi, Efstratios Mylonas, Dina Kotsifaki, Mary Providaki, Charalambos Pozidis, Melina Fotiadou, and Michael Kokkinidis. Probing protein folding with sequence-reversed α-helical bundles. *International journal of molecular sciences*, 22(4): 1955, 2021.