Structural representations of DNA regulatory substrates can enhance sequence-based algorithms by associating functional sequence variants

Jan Zrimec
janzrimec@gmail.com
Chalmers University of Technology
Gothenburg, Sweden

ABSTRACT

The nucleotide sequence representation of DNA can be inadequate for resolving protein-DNA binding sites and regulatory substrates, such as those involved in gene expression and horizontal gene transfer. Considering that sequence-like representations are algorithmically very useful, here we fused over 60 currently available DNA physicochemical and conformational variables into compact structural representations that can encode single DNA binding sites to whole regulatory regions. We find that the main structural components reflect key properties of protein-DNA interactions and can be condensed to the amount of information found in a single nucleotide position. The most accurate structural representations compress functional DNA sequence variants by 30% to 50%, as each instance encodes from tens to thousands of sequences. We show that a structural distance function discriminates among groups of DNA substrates more accurately than nucleotide sequence-based metrics. As this opens up a variety of implementation possibilities, we develop and test a distance-based alignment algorithm, demonstrating the potential of using the structural representations to enhance sequence-based algorithms. Due to the bias of most current bioinformatic methods to nucleotide sequence representations, it is possible that considerable performance increases might still be achievable with such solutions.

CCS CONCEPTS

• **Applied computing** → **Bioinformatics**.

KEYWORDS

Regulatory genomics, DNA structural properties, Bio-algorithms

ACM Reference Format:
Jan Zrimec. 2020. Structural representations of DNA regulatory substrates can enhance sequence-based algorithms by associating functional sequence variants. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '20), September 21–24, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3388440.3412482

BCB '20, September 21–24, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7964-9/20/09…$15.00
https://doi.org/10.1145/3388440.3412482

1 INTRODUCTION

Besides the popular yet simplistic representation of DNA as a polymer chain of 4 different nucleotide bases, the molecule in its double stranded form possesses certain conformational and physicochemical properties. These properties are important for interactions of the DNA with proteins, which drive many essential cellular processes [10, 15, 19, 20]. These include but are not limited to transcription [10], replication [3] as well as horizontal gene transfer (HGT) [16, 20]. The processes are commonly initiated and regulated at specific DNA regulatory regions, which are the substrates for protein binding and enzymatic activity, and include promoters [15], origins of replication [3] and origins of transfer (transfer regions, plasmid conjugation in HGT) [4]. The DNA substrates contain either one or multiple protein binding sites, such as, e.g., transcription factor binding sites (TFBS) in promoters, as well as additional sequence and structure context that is related to protein-DNA recognition and binding preceding the main enzymatic processing [8, 9, 20].

As such, the nucleotide sequence representation is often not sufficiently informative for discriminating DNA regulatory substrates, as it encodes specific conserved structural properties that are not directly obvious from the mere sequence context (Table S1 on Github) [8, 14, 20]. To uncover the encoded structural properties, many DNA structure models and prediction tools have been developed, including (i) DNA thermodynamic stability and its potential for destabilization and melting bubble formation [19], (ii) DNA major and minor groove properties that describe their accessibility by proteins [2, 10], (iii) DNA intrinsic curvature [5] and flexibility [15], (vi) DNA twisting and supercoiling [15], (v) differences in DNA spacing and orientation between binding and enzymatic sites [5], and (vi) the propensity for transitions between DNA forms, such as from B-DNA to A-DNA or to Z-DNA [7, 15]. However, due to the large variety and differences among the DNA structural models, it is not simple to choose among them or integrate them within existing DNA bioinformatic frameworks. Current studies thus focus merely on specific groups of DNA structural properties [3, 10, 11, 18, 20], whereas a common DNA structural representation spanning the whole structural repertoire could help to improve existing frameworks and facilitate the development of new DNA algorithms.

The aim of the present study was to analyse and engineer novel DNA structural representations for use with bioinformatic frameworks, such as sequence alignment and motif finding algorithms, and to compare them with the standard nucleotide sequence-based methods. First, we construct different DNA structural representations by applying dimensionality reduction techniques to over 60

DNA properties involved in DNA-protein interactions. To explore how much information can sufficiently describe functional DNA regions, we compress the representations using clustering algorithms to an estimated 2 to 8 bits. Next, we explore the capability of each representation to encode multiple DNA sequence variants using TFBS motif datasets. To enable the use of the representations with existing DNA algorithms, by comparing the similarity of encoded sequences, we develop a structural distance function and test it on datasets of transfer and promoter regions. Finally, we test the representations within a sequence alignment framework and discuss ideas for hybrid sequence and structure-based approaches for analysing regulatory DNA substrates.

2 METHODS

2.1 Datasets

We obtained 64 published DNA structure models [16] based on nearest neighbor dinucleotide (56), trinucleotide (4) and pentanucleotide (4) models. They comprised physicochemical and conformational properties and properties attributed to DNA-protein interactions, and included the widely used models DNA shape [10], Orchid [2], DNA stability and duplex destabilization [19].

A dataset (sites file) of transcription factor binding site (TFBS) motifs was obtained from the Jaspar database (jaspar.genereg.net) and filtered to contain only sequences with {A,C,G,T} characters of equal length as the median length in each motif group.

A published dataset of transfer regions from 4 mobility (Mob) groups [20] was used as positive examples and expanded with 64 negative examples. Negative example sequences were selected randomly from a region 200 to 800 bp around the enzymatic nicking sites [20], thus containing different non-regulatory coding and non-coding regions, and low sequence similarity was verified among the sequences (p-distance > 0.6). The part of the transfer regions with relevant protein binding features from -140 bp to +80 bp according to the nicking site was used [20]. For testing the alignment algorithm, a second published dataset of 112 transfer regions (queries) of equal length as the ones above and 52 plasmid targets from 4 Mob groups was used [16].

A dataset of Escherichia coli promoter regions was obtained [6] with 100 bp positive and negative examples (positive, mixed1 and control). We randomly sampled 200 sequences from each dataset to create a 600 element dataset. The second dataset of Escherichia coli promoter regions was obtained from Regulon DB v9 (regulondb.ccg.unam.mx) and contained 81 bp positive examples grouped according to 6 sigma factors [15]. A random sampling of 94 sequences (size of smallest group) from each group yielded a dataset with 564 elements.

2.2 Construction and analysis of DNA structural representations

To develop DNA structural representations we computed the structural properties of all permutations of k-mers 3, 5, 7 and 9 bp in length (1 to 4 neighboring regions around a specific nucleotide), performed dimensionality reduction and clustering (Fig. 1A). The structurally defined groups of k-mers were termed s-mers. Structural properties were calculated in windows of 5 bp or at default values as described [19, 20]. Dimensionality reduction and analysis of the main components of variance was performed using Principal Component Analysis (PCA). The k-means clustering algorithm was used (Matlab), where clusters with a lowest total sum of distances were chosen from 10 runs of up to 1000 iterations at default settings. The optimal amount of clusters was analysed with the Elbow, average Silhouette and GAP methods with Matlab function evalclusters at default settings with triplicate runs. The tested numbers of clusters included 4, 8, 16, 32, 64, 128 and 256 clusters, chosen considering that (i) positions with 2^x possible states (clusters) can carry a maximum of x bits of information [15], (ii) 1 bp of DNA carries up to 2 bits of information (iii) up to 4 bp neighboring regions defined the structural effect in the s-mers, and (iv) s-mers are overlapping, meaning information is distributed among all of them. Thus, up to 8 bits of information (256 clusters) was expected per s-mer (and less with decreasing s-mer size).

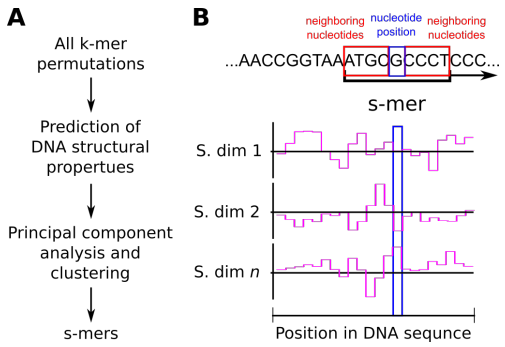

Figure 1: Schematic depiction of the (A) construction and (B) usage of structural representations. A structural representation of a given DNA sequence, where each central nucleotide position and its neighboring regions define a k-mer of 3-9 bp and are encoded as an s-mer with *n* structural dimensions (S. dim.), is defined as a sequence of s-mer cluster centroids.

For a DNA substrate, the length of the structural representation was equal to the length of the nucleotide sequence minus the leftover nucleotides at the borders equal to $(s-1)/2$, due to the neighboring nucleotides in s-mers (Fig. 1B). The s-distance between two DNA substrates was the sum of squared Euclidean distances between the cluster centroids of all equally positioned s-mers in their structural representations of length n,

$$s-distance = \sum_{i=1}^{n} d(C_{1i}, C_{2i})^2, \quad (1)$$

where $C_{ni} = (c_{n1}, c_{n2}, ..., c_{nk})$ are the cluster centroids of the s-mer at position i of the first and second sequences, respectively. The p-distance was equal to the Hamming distance corrected for sequence length.

2.3 Statistical analysis and performance metrics

For statistical hypothesis testing, the Python package Scipy v1.1.0 was used with default settings. To evaluate the explained variation,

the coefficient of determination was defined as

$$R^2 = 1 - \frac{SS_{Residual}}{SS_{Total}}, \quad (2)$$

where $SS_{Residual}$ is the within group residual sum of squares and SS_{Total} is the total sum of squares. The compression ratio was calculated as

$$Compression\ ratio = \frac{\text{Num. unique k-mers}}{\text{Num. unique s-mers}}. \quad (3)$$

Precision and recall were defined as

$$Precision = \frac{TP}{TP + FP},\ Recall = \frac{TP}{TP + FN}, \quad (4)$$

where *TP*, *FP* and *FN* denote the number of true positive, false positive and false negative elements, respectively. The F_1-score was defined as

$$F_1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall}. \quad (5)$$

The *F*-test was performed using permutational multivariate analysis of variance with sequence bootstraps [1, 20]. The distribution of the *F* function under the null hypothesis of no differences among group means was evaluated by performing 1e4 bootstrap repetitions, with *p*-values calculated as

$$p = \frac{\text{Num. } F_{Bootstrap} \geq F}{\text{Total num. } F_{Bootstrap}}. \quad (6)$$

2.4 Alignment algorithm

We developed and tested a simple ungapped DNA sequence alignment framework (Algorithm 1) that finds the most similar segments to query sequences in target sequences using a given distance function. The assessment of algorithm performance included (i) locating the transfer regions to within +/- 1 bp of their known locations in the target plasmids and (ii) correct typing of Mob groups in the target plasmids. For this, true and false positive and negative counts were obtained from the alignment tests by considering only the lowest scoring hit per alignment. A true or false positive value was assigned if the result was below a specified significance cutoff and corresponded or not, respectively, to the known value (region location or Mob group), and alternatively, a false or true negative value was assigned to results above the significance cutoff that corresponded or not, respectively, to the known value. Statistical significance of distance scores (at *p*-value cutoffs from 1e-6 to 1e-1) was evaluated using bootstrap permutations (n = 1e6 per sequence) of 10 randomly selected query sequences (Eq. 6) and a mapping function between the distance scores and *p*-values was obtained by least squares curve fitting (Matlab) to a second order polynomial function (distance score of 0 corresponded to the theoretical limit of 1e-132) [16].

Algorithm 1: Sequence alignment algorithm.

```
input query_set, target_set;
for i = 1 : size(query_set):
  for j = 1 : size(target_set):
    for k = 1 : length(target_set(j)):
      dist(i,j,k) = distance(query_set(i),target_set(j)(k));
return min(dist(:,:)).
```

Table 1: Summary of s-mer construction. Groups of s-mers of size *s* comprised all permutations of k-mers of 3-9 bp, with *n* neighboring nucleotides (Fig. 1B). The amount of structural models used (Struct.) varied due to sequence length constraints. Each structural representation used principal components (PC) describing over 99% of data variance.

s	*n*	Perm.	Struct.	PC>0.99	Clust. size *k*
3	1	64	57	14	2^2 to 2^6
5	2	1024	62	17	2^2 to 2^8
7	3	16384	64	18	2^2 to 2^8
9	4	262144	64	18	2^2 to 2^8

2.5 Software

Matlab v2017 (www.mathworks.com), Python v3.6 (www.python.org) and R v3.5 (www.r-project.org) were used. Code and supplements are available at www.github.com/JanZrimec/smer_acm_bcb_20.

3 RESULTS

3.1 Fusion of DNA structural properties into a compact representation encapsulates main protein-DNA binding features

Considering that the first 4 neighboring nucleotides have the largest effect on the structural state of a given nucleotide base pair [19], we designed nucleotide-position specific DNA structural representations that included the effects of 1 to 4 neighboring nucleotides, termed s-mers (Table 1, Fig. 1, Methods 1). The s-mers were based on calculating up to 64 of the most widely used DNA structural properties for all permutations of the corresponding equally sized k-mers, followed by dimensionality reduction and clustering (Fig. 1A, Methods 2). The calculated DNA structural properties included experimental physicochemical, conformational and protein-DNA binding variables [16].

Dimensionality reduction with Principal Component Analysis (PCA) yielded a specific number of principal components (PC) with each s-mer size *s* that explained over 99% of the data variance (Fig. 2A). The amount of PCs increased with the s-mer size and with the number of initial structural variables from 14 to 18 PCs, which was an over 3.5-fold decrease in the amount of variables required to describe almost all of the original information (Table 1). On average, 3, 6, 9 and 17 components were required to explain over 60, 80, 90, and 99% of the variance, respectively. The coefficient of variation (σ/μ) of the first 6 most informative PCs across the s-mer sizes was below 0.637 and lowest with the first and sixth components with 0.160 and 0.466, respectively, showing that these PCs carried similar structural information with each *s*. Analysis of PCA loadings showed that each PC mainly comprised a number of distinct structural variables, which enabled us to determine the key protein-DNA binding features defined by the 6 most informative components (in the order of decreasing importance): thermodynamic stability, horizontal flexibility, torsional flexibility, conformational stability, major and minor groove accessibility and A-DNA to B-DNA transition potential. Moreover, the top 20 sorted structural variables

contained 3 of the well known DNAshape functions [10] and OR-ChID2 [2], with nucleosome positioning (phase) [12] being the most important.

Although the dimensionality reduced structural data was not expected to form strong clusters, a limit must exist to the resolution of the structural representations, above which there is no measurable influence on the achievable computational accuracy, and an additional level of compactness of the DNA codes was desirable. Clustering was thus performed with the number of clusters k varied between 4 and 256 clusters (2 to 8 bits), with the exception of using up to 32 clusters with 3-mers (Table 1, Methods 2). Standard cluster evaluation methods, including Elbow and Silhouette, showed that with a decreasing k, the overall accuracy of the clustered data representations decreased compared to using a higher number of clusters. At the highest k (256), the explained variance (R^2) was over 80% with both s-mer sizes 5 and 7 (s = 9 not fully tested due to memory restrictions). With a decreasing number of clusters, progressively larger clusters were obtained (Fig. 2B) with a decrease in the percentage of explained variance down to 40% with k = 4 clusters, and similarly a decrease in the average Silhouette ratio. The cluster sizes were approximately normally distributed (Fig. 2B), with the variation of cluster sizes increasing with an increasing k, although the coefficient of variation did not surpass 0.52.

3.2 Structural representations encode groups of functional sequence variants

We next explored whether the s-mers could encode groups of conserved functional DNA motifs [8], and if the new representations could encapsulate DNA motifs more compactly than bare k-mers. We used a dataset of 595 Jaspar transcription factor binding site (TFBS) motifs comprising 1,296,654 unique DNA sequences from multiple model organisms (Methods 1), which contained at least 10 motifs per group and were from 9 to 29 bp long (lower limit set by largest s-mer size). To test the capacity of the structural representations to encode TFBS motifs we first measured the *compression ratio* of the amount of unique s-mers versus the amount of unique k-mers observed with a given motif (Eq. 3). Due to computational complexity and memory limitations, only certain parameter combinations could be tested. We observed an increase in the average *compression ratio* across motifs with an increasing s-mer size (Fig. 2C), as it increased from 1.132 with s = 5 (k = 256) to 1.296 with s = 9 (k = 256). Similarly, the average compression ratio increased with decreasing amount of clusters (Fig. 2C), from 1.132 with k = 256 (s = 5) to 3.152 with k = 4 (s = 3). The variation of the compression ratio across the different motifs was approximately constant (SD between 0.056 with s = 3, k = 4 and 0.092 with s = 7, k = 128). Moreover, we measured a significant negative correlation (Pearson's r, p-value < 4.7e-5) between the *compression ratio* and the TFBS sequence length increasing from -0.166 to -0.667 with an increasing s-mer size s = 3 (k = 32) to s = 9 (k = 128), respectively, and up to -0.724 with a decreasing cluster size k (s = 3, k = 4). Weak correlation (Pearson's r, p-value < 0.026) was also observed between the *compression ratio* and the number of unique sequences in a TFBS motif, similarly as above increasing from -0.091 to -0.311 with an increasing s = 3 (k = 32) to s = 9 (k = 128), respectively, and up to -0.382 with a decreasing k (s = 3, k = 4). This suggested that the capacity for compression decreases with more abundant DNA sequence space, such as with longer motifs or ones with a more diverse set of sequence variants.

Since the structural representations indeed compressed the TFBS motifs up to 3-fold, we next explored how accurate the encodings were at describing different functional motif variants. We selected the most sequence-abundant motif, the 18 bp Human MAFF motif (Jaspar: MA0495.1, class of Basic leucine zipper factors) that contained 49,462 unique sequence variants. Using a randomly selected 1% (n = 495) subset of these sequence variants to define their s-mers, we measured how many of the remaining 99% of the motifs were described by (or rather, could be predicted from) these structural encodings. This meant reconstructing all the possible motif sequence variants from each structural representation instance, and gave an estimate of the encoding accuracy. The initial *precision* and *recall* (Methods 3) obtained without any encoding were 1 and 0.01, respectively. Unsurprisingly, an inverse relation was observed between *precision* and *recall* (Fig. 2D). *Precision* was highest (0.838) with a low s-mer size (s = 3, k = 32) and decreased 6.5-fold with an increasing s (0.129, s = 9, k = 256), whereas *recall* increased by 9% from 0.0102 to 0.0111 at the equal parameter values, respectively. A reason for the decrease in *precision* was likely that the average amount of distinct sequence variants encapsulated by the different structural representations increased with an increasing s-mer size as well as with a decreasing k (Fig. 2E). This related to an increasing coverage of the correct TFBS motifs (true positives) as well as an increasing number of variants not in the given TFBS sequence set (false positives). However, it is also possible that the TFBS sequence set is incomplete and does not contain all the possible functional sequence variants of that motif, meaning that the true positive and negative space is unknown. With the 18 bp MAFF motif, we estimated that up to 6.9e10 sequence variants could exist, whereas the set of sequence variants in the TFBS represented merely 7.2e-7 of this diversity and was likely undersampled.

3.3 Structure-based distance function resolves regulatory DNA more accurately than sequence-based ones

To facilitate the comparison of two different structural representations (Fig. 1: s-mer vectors), such as is done with nucleotide sequences using e.g. the p-distance, we defined the structural s-distance as the sum of squared Euclidean distances between s-mer cluster centroids of two representations (Eq. 1). To test the s-distance and compare it to the p-distance (Methods 2), we used a dataset of whole DNA regulatory regions that control the initiation of DNA transfer in plasmid conjugation [4] (Methods 1). These 220 bp transfer regions comprise binding sites for enzymes and accessory proteins that regulate transfer. The dataset contained 64 transfer regions [20] from 4 mobility groups (Mob, defined by amino acid homology of the main transfer enzyme) [4], each with an approx. equal amount of 16 elements. In order to additionally test the possibility of discriminating between functional and non-functional transfer sequences, here we expanded the dataset to include, besides the positive sequences (Pos) also an equal amount negative counterparts (Neg, Methods 1). Using the specific distance functions as the measure of variation across the data within a permutational MANOVA framework (bootstrap n = 1e4, Methods 3)

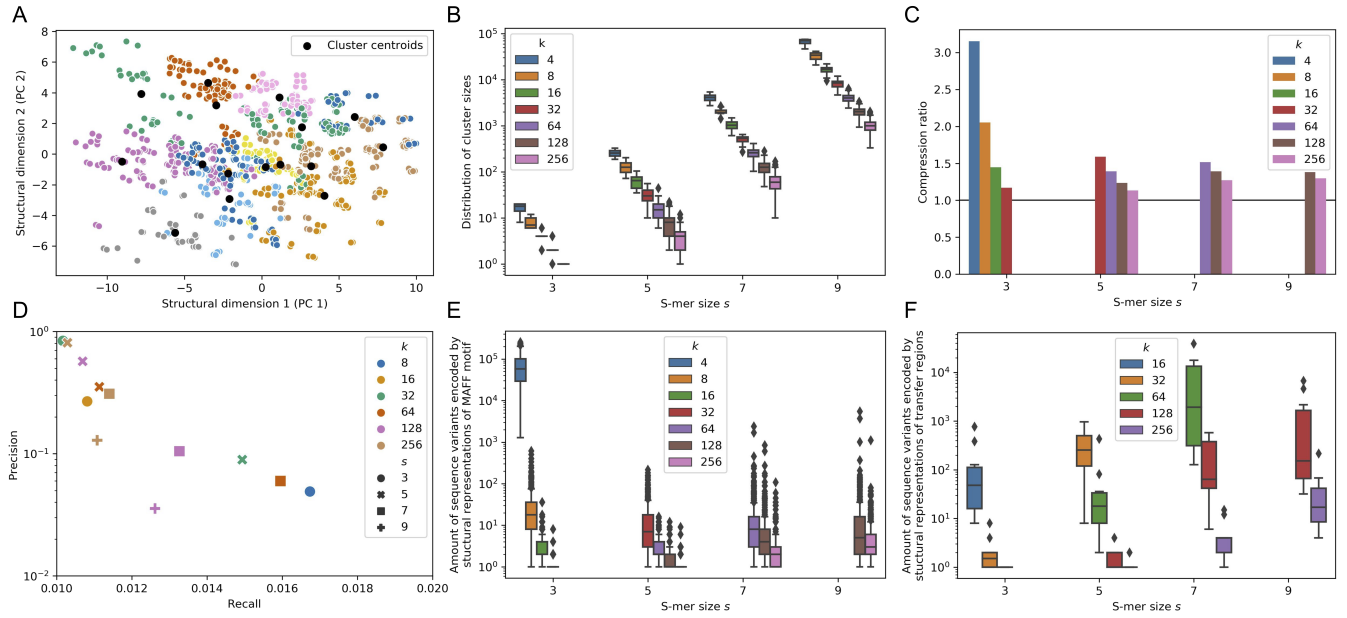

Figure 2: (A) First two principal components (PC) of the structural representation with s-mer size $s = 5$ and num. clusters $k = 16$, where each point represents a different 5-mer nucleotide permutation. Colors depict the clusters and black points denote the cluster centroids. (B) Distributions of cluster sizes across the structural representations with different s and k. (C) Compression ratios of Jaspar TFBS motifs obtained with different structural representations, black line denotes no compression. (D) Precision vs. recall when comparing an encoding of 1% of sequence variants of the Human MAFF motif (Jaspar: MA0495.1) with the remaining 99% of sequence variants, at different parameters s and k. Precision and recall obtained with nucleotide k-mers were 1 and 0.01, respectively. (E) Amount of unique DNA sequence variants encoded by structural representations of the 18 bp MAFF motif, results with 1% ($n = 100$) of motif variants shown. (F) Amount of unique DNA sequence variants encoded by structural representations of 220 bp plasmid transfer regions, results with 10 region variants shown.

[1], we observed significant (p-value $\leq$ 1e-4) discrimination of both the Pos/Neg examples as well as MOB groups with the s-distance, which was not the case with the p-distance (Table 2). On average, when discriminating Pos/Neg examples and MOB groups with the s-distance, the amount of explained variance (R^2, Eq. 2) increased 3-fold and 2.2-fold, respectively, compared to using the p-distance (Table 2). Additionally, using two datasets of Escherichia coli promoter regions, one with positive and negative examples [6] and the other grouped according to 6 most prevalent sigma factors [15] (Methods 1), we verified the above results, as significant (p-value < 0.05) group discrimination could be achieved with structural representations and not with the p-distance.

Furthermore, to determine how the amount of encoded sequence variants scales with the length of the sequence, we computed the amount of sequence variants encoded by structural representations of the 220 bp transfer regions, by randomly selecting 10 transfer regions (Fig. 2F). Compared to the 18 bp TFBS motifs, structural representations of the 220 bp transfer regions encoded, on average, an over 32-fold higher number of sequence variants (Fig. 2F).

3.4 DNA sequence-based algorithms can be enhanced with structural representations

We tested whether the structural representations and s-distance could be applied to existing algorithm frameworks, such as sequence alignment. The alignment framework that we developed (Algorithm 1) enabled the use of different metrics such as the s-distance and could align a query and a target sequence based on finding the minimal distance between them. Since equal sized groups were no longer required, we used an expanded dataset of 112 transfer regions as the query dataset and a target dataset of 52 plasmids with known Mob groups and locations of the transfer regions [16] (Methods 1). By counting the amount of lowest-distance alignments in the target dataset that were below a specified distance threshold, we could define similar metrics as for a binary classification problem, namely amounts of true and false positive and negative results (Methods 4) as well as the harmonic mean of precision and recall, the F_1-score (Eq. 5). Although both distance functions proved accurate, we observed, on average, an over 7% improvement of the F_1-score with the s-distance compared to the p-distance, both when discriminating functional regions (Pos/Neg) as well as the Mob groups (p-value < 1e-13, Table 2). The s-distance based algorithm (at $s = 7$, $k = 128$) thus correctly uncovered 32 (62%) transfer regions in the target set with 30 of them (58%) correctly Mob typed, compared to 29 (56%) and 26 (50%) with the sequence based p-distance, respectively. This suggested that existing DNA sequence-based algorithms could indeed be enhanced with structural representations [8–10] (Fig. S1 on Github).

Table 2: Comparison of nucleotide p-distance and structural s-distance functions, at parameters s-mer size *s* and num. clusters *k*. 'P/N' denotes discrimination of Pos/Neg examples, F_1-scores were obtained with alignment algorithm (1).

Func.	*s*, *k*	R^2		ANOVA *p*-val		F_1-score	
		P/N	Mob	P/N	Mob	P/N	Mob
p-dist.	/	0.017	0.113	0.472	0.473	0.892	0.834
s-dist.	3, 32	0.046	0.237	< 1e-4	< 1e-4	0.955	0.853
	5, 128	0.049	0.248	< 1e-4	< 1e-4	0.942	0.923
	7, 128	0.054	0.253	1e-4	< 1e-4	0.970	0.923
	9, 128	0.057	0.264	1e-4	< 1e-4	0.954	0.879

4 DISCUSSION

Here we used 64 functionally-relevant DNA structural properties (Table 1) and fused them into compact DNA structural representations containing the most important structural information (e.g. 6 components carried over 80% of the initial data variance). Recovery of the key distinguishing properties of the first 6 structural components showed that they indeed reflected the main properties involved in protein-DNA interactions (Table S1 on Github). The amount of structural information could be further refined with clustering down to 2 bits (Fig. 2B), where the compression ratio of functional DNA sequence motifs could be as high as 3:1 (Fig. 2C). Nevertheless, the most promising results were obtained with a number of clusters corresponding to 6 to 8 bits of information (64 to 256 clusters, Fig. 2D). These structural representations could compress functional sequence variants by around 30% to 50%, where each instance of the structural representation encapsulated up to 20 sequence variants of a TFBS motif (Fig. 2E) and up to 2000 variants of a whole DNA regulatory region (Fig. 2F). Further testing is required, however, using datasets with a more complete coverage of the functional sequence space or experimentally, to properly investigate the capacity to encode groups of functional DNA sequence variants and their conserved functional properties.

Devising the s-distance function opens up a plethora of possibilities for testing the structural codes and implementing them into DNA algorithms, as it was found to resolve regulatory DNA more accurately than sequence-based metrics [16, 20] (Table 2). One can also think of additional metrics that can prove useful, such as for instance a Jaccard distance using s-mers instead of k-mers [16], that can distinguish coding and non-coding sequences across genomes or help bin species in metagenomic data. Furthermore, by developing and testing a distance-based alignment algorithm (1), we demonstrated the potential of the structural representations to enhance existing sequence-based algorithms. Here, the usefulness of sequence-like codes (similar to the nucleotide code ACGT) stood out, as the structural representations could be realized as mere sequences of cluster indices with precomputed pairwise distances, abstracting from, and altogether disposing of, the structural component space. This can simplify their implementation in existing sequence-based algorithms and also improves the accuracy of pinpointing enzymatic sites, such as nicking sites in transfer regions, down to a resolution of 1 bp [16]. Accordingly, the solution can uncover many new variants of transfer regions in natural plasmids, helping researchers investigate the potential for plasmid mobility and its global effects [16].

In contrast to machine learning (ML), where models learn to use only specific subsets of the structural variables for different tasks, such as discriminating Pos/Neg examples or MOB groups [18–20], DNA sequence analysis frequently requires generalizing across multiple tasks and using all variables. We have found that sequence based models of structural properties indeed outperform ML models [16, 20], except with deep learning, which changes the paradigm by being able to infer new data representations from mere nucleotide sequence [17].

Besides sequence alignment [16], the potential uses of the structural representations include: (i) phylogenetic analysis of regulatory DNA regions [4], (ii) analysis of single nucleotide variations [15], as the structural representations contain variants with position-specific nucleotide substitutions, (iii) motif identification [11], where, for instance, the initialization stage of graph-based algorithms could be performed using structural representations [13], and (iv) design of DNA substrates with a modified DNA sequence but conserved functionality. Other, combined approaches could potentially mimic the protein-DNA search and binding dynamics in DNA regulatory regions [8–10] (Fig. S1 on Github) and adopt a combination of both structural and sequence features, or use different representations for the non-coding (structural) and coding (sequence) regions for instance in gene identification.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency [n° Z2-7257]. I kindly thank Tomaz Pisanski, Filip Buric, Ziva Stepancic, Chrats Melkonian, Maria Pilar Garcillán-Barcia, Fernando de la Cruz, Joshua Ramsey and Ales Lapanje for discussions and support.

REFERENCES

[1] Anderson M J 2001 *Austral Ecol.* **26**(1), 32–46.
[2] Bishop E P, Rohs R, Parker S C J, West S M, Liu P, Mann R S, Honig B and Tullius T D 2011 *ACS Chem. Biol.* **6**(12), 1314–1320.
[3] Chen W, Feng P and Lin H 2012 *FEBS Lett.* **586**(6), 934–938.
[4] Garcillán-Barcia M P, Francia M V and de la Cruz F 2009 *FEMS Microbiol. Rev.* **33**(3), 657–687.
[5] Geggier S and Vologodskii A 2010 *Proc. Natl. Acad. Sci. U. S. A.* **107**(35), 15421–15426.
[6] Gusmão E G and de Souto M C P 2014 *in* '2014 International Joint Conference on Neural Networks (IJCNN)' IEEE pp. 494–501.
[7] Kulkarni M and Mukherjee A 2013 *J. Chem. Phys.* **139**(15), 155102.
[8] Levo M, Zalckvar E, Sharon E, Dantas Machado A C, Kalma Y, Lotam-Pompan M, Weinberger A, Yakhini Z, Rohs R and Segal E 2015 *Genome Res.* **25**(7), 1018–1029.
[9] Marcovitz A and Levy Y 2013 *The Journal of Physical Chemistry B* **117**(42), 13005–13014.
[10] Rohs R, West S M, Sosinsky A, Liu P, Mann R S and Honig B 2009 *Nature* **461**(7268), 1248–1253.
[11] Samee M A H, Bruneau B G and Pollard K S 2019 *Cell Systems* **8**(1), 27–42.e6.
[12] Satchwell S C, Drew H R and Travers A A 1986 *J. Mol. Biol.* **191**(4), 659–675.
[13] Stepančič Ž 2014 *Journal of Computational Biology* **21**(10), 741–752.
[14] Tosato V, West N, Zrimec J, Nikitin D V, Del Sal G, Marano R, Breitenbach M and Bruschi C V 2017 *Front. Oncol.* **7**, 231.
[15] Watson J D, Baker T A, Bell S P, Gann A, Levine M K and Losick R 2008 *Molecular Biology of the Gene. 6th. ed* Pearson/Benjamin Cummings.
[16] Zrimec J 2020 *bioRxiv* p. 2020.04.20.050401.
[17] Zrimec J, Buric F, Muhammad A S, Chen R, Verendel V and others 2019 *bioRxiv* p. 10.1101/792531.
[18] Zrimec J, Kopinč R, Rijavec T, Zrimec T and Lapanje A 2013 *J. Microbiol. Methods* **95**(2), 186–194.
[19] Zrimec J and Lapanje A 2015 *IEEE/ACM Trans. Comput. Biol. Bioinform.* **12**(5), 1137–1145.
[20] Zrimec J and Lapanje A 2018 *Sci. Rep.* **8**(1), 1820.