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

Jan Zrimec
janzrimec@gmail.com
Department of Biology and Biological Engineering
Chalmers University of Technology
Gothenburg, Sweden

ABSTRACT

The nucleotide sequence representation of DNA can be inadequate for resolving protein-DNA binding sites and DNA 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 protein-DNA binding sites as well as whole regulatory substrates. We show that the main structural components reflect key properties of protein-DNA interactions and further compress the representations down to the amount of information found in a single nucleotide. The most accurate structural representations compress functional sequence variants by 30% to 50%, as each instance encodes from tens to thousands of functional DNA sequence variants. 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 a considerable performance increase might still be achievable with such novel solutions.

CCS CONCEPTS

• **Applied computing** → **Molecular sequence analysis**; **Recognition of genes and regulatory elements**; **Bioinformatics**.

KEYWORDS

Bioinformatic algorithms, DNA structural properties, Transcriptional regulation, Horizontal gene transfer

1 INTRODUCTION

Besides the popular yet simplistic representation of DNA as a polymer chain of 4 different nucleotide bases A, C, G and T, the molecule in its double stranded dsDNA form possesses certain conformational and physicochemical properties. These are especially important in relation to interactions of the DNA with proteins, which drive many essential cellular processes [1–3]. These include but are not limited to transcription [1], replication [4] as well as horizontal gene transfer (HGT) [3, 5]. 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 [6], origins of replication [4] and origins of transfer (transfer regions, plasmid conjugation in HGT) [7]. The DNA substrates contain either one or multiple protein binding sites, such as for instance 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 [3, 8, 9].

As such, the nucleotide sequence representation is often not sufficiently detailed for discriminating DNA regulatory substrates, as it encodes specific conserved structural properties that are not directly obvious from the mere sequence context (Table 1) [3, 10, 11]. 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, modelled with the nearest neighbor (NN) framework [12] and thermally induced duplex destabilization [2], (ii) DNA major and minor groove properties that describe their size and thus accessibility by proteins, which are the focus of the DNAshape and ORChID2 models [13, 14] among others, (iii) DNA intrinsic curvature and flexibility, modeled by measuring DNAzeI digestion profiles [15] and DNA persistence lengths [16], (vi) DNA twisting and supercoiling, captured by multiple DNA conformational models and variables [17–19], (v) differences in DNA spacing and orientation between binding and enzymatic sites, described for instance by DNA helical repeats [16], and (vi) the propensity for transitions between DNA forms, such as from B-DNA to A-DNA or to Z-DNA given with respective variables [20–23]. However, due to the large variety and differences among the DNA structural models and variables, it is not simple to choose among them or integrate them within existing DNA bioinformatic frameworks.

Current studies thus focus on specific groups of DNA structural properties and do not span the whole possible structural repertoire [3, 4, 14, 24–26]. A common DNA structural representation could thus help circumvent these problems and facilitate the development of novel, improved DNA algorithms.

Hence, 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, based on DNA physicochemical and conformational properties that are involved in DNA-protein interactions and using dimensionality reduction techniques we constructed different structural DNA representations. To explore the smallest amount of structural information that could sufficiently describe functional DNA regions we compressed the DNA representations using clustering algorithms to encompass from an estimated 2 to 8 bits of information. Next, we explored the capability of each representation to encode multiple DNA sequence variants using TFBS motif datasets. To facilitate the usefulness of the representations with existing DNA algorithms via comparison of the (dis)similarity of encoded sequences, we developed a structural distance function and tested it on datasets of transfer and promoter regions. Finally, we explored the applicability of the structural representations within a sequence alignment framework and discussed ideas for hybrid sequence and structure-based approaches specifically for analysing regulatory DNA substrates.

2 METHODS

2.1 Datasets

We obtained 64 published DNA structure models that were shown to be informative for analysis of DNA-protein interactions [5] (Table 1). These models were based on nearest neighbor dinucleotide (56), trinucleotide (4) and pentanucleotide (4) models and included physicochemical and conformational properties and properties attributed to DNA-protein interactions. Of these, 44 models were derived experimentally and the rest computationally based on experimental data. The set included the widely used models DNA shape [1, 14], Orchid [13], DNA stability and thermally induced duplex destabilization [2, 12].

A dataset of transcription factor binding site (TFBS) motifs was obtained from the Jaspar database [31] (sites file) and filtered to contain only sequences with {A,C,G,T} characters, of equal length as the median length in each motif group but at least 9 bp long and containing at least 10 motif sequence variants.

A published dataset of transfer regions from 4 mobility (Mob) groups [3] 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 [3], 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 [3]. Sequences below a 15% sequence identity (33 bp) were merged (elements RSF1010 and R1162, pSymA and pSymB). For testing the alignment algorithm, a second published dataset of 112 transfer regions (queries) of equal length as

the ones above and 52 plasmids (targets) [5] from 4 Mob groups was used.

A dataset of Escherichia coli promoter regions was obtained [32] 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 [33] and contained only 81 bp positive examples grouped according to 6 sigma factors [6]. 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) and performed dimensionality reduction followed by 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 [2, 3]. 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 [34] and GAP [35] 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 [36], (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).

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,

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

where $C_{1i} = (c_{11}, c_{12}, ..., c_{1k})$ is the cluster centroid of the s-mer at position i of the first sequence and $C_{2i} = (c_{21}, c_{22}, ..., c_{2k})$ is the cluster centroid of s-mer at position i of the second sequence.

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

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

Table 1: Overview of DNA structural properties and representative variables in protein-DNA interaction.

| DNA structural properties	Facilitated protein-DNA interactions	Representative structural variables	References
DNA stability and propensity for destabilizations and melting bubble formation	Enzymatic processing of substrates, e.g. relaxase nicking of transfer regions leads to secondary structure formation	Duplex stability, Thermally induced duplex destabilization (TIDD)	[2, 12, 27, 28]
Major and minor groove properties	Readout of chemical information, e.g. transcription factors in promoters	DNAShape, ORChID2	[1, 6, 13, 14]
Intrinsically curved or flexible regions	Binding and topological changes, e.g. IHF binding in promoters	DNAzeI cleavage frequency, Persistence length	[15, 16, 29]
DNA twist and supercoiling	Topological changes recognized by proteins, e.g. histones, and affect multiple other properties	Twist and other conformational variables	[6, 17–19]
Differences in DNA spacing and orientation in binding and enzymatic sites	Affect binding with multiple contact points and protein complex formation	Helical repeats	[6, 16, 30]
Propensity for transitions between DNA forms B-DNA, A-DNA, Z-DNA	Affect overall features recognized by proteins and their accessibility	B-A and B-Z transition propensities	[20–23]
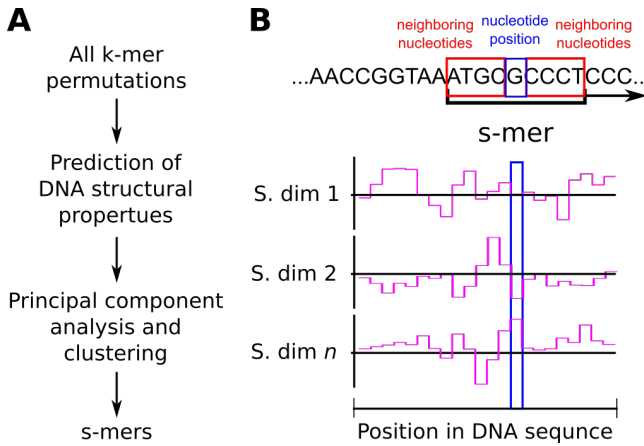
Figure 1: Schematic depiction of the (A) construction and (B) usage of structural representations. In a structural representation of a given DNA sequence, each central nucleotide position and its neighboring regions define a k-mer from 3 to 9 bp in length, and are encoded as an s-mer with n structural dimensions (S. dim.) that can be defined as a sequence of s-mer cluster centroids.

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

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

$$Recall = \frac{TP}{TP + FN}, \quad (5)$$

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

$$F_1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall}. \quad (6)$$

The F-test was performed using permutational multivariate analysis of variance with sequence bootstraps [3, 37]. 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

$$p = \frac{\text{Num. } F_{Bootstrap} \geq F}{\text{Total num. } F_{Bootstrap}}. \quad (7)$$

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 did not correspond, 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 did not correspond, respectively, to the known value. The 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. 7) and a mapping function between the distance scores and p-values was then 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) [5].

Algorithm 1: Sequence alignment algorithm.

---

input query_set, target_set;
for *i = 1 : size(query_set)* **do**
    for *j = 1 : size(target_set)* **do**
        for *k = 1 : length(target_set(j))* **do**
            dist(i,j,k) = distance(query_set(i),target_set(j)(k));

        end
    end
end
return min(dist(:,:)).

2.5 Software

Matlab v2017 (www.mathworks.com), Python v3.6 (www.python.org) and R v3.5 (www.r-project.org) were used. The code is 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 [2, 38], we designed nucleotide-position specific DNA structural representations that included the effects of 1 to 4 neighboring nucleotides, termed s-mers (Table 2: sizes 3, 5, 7 and 9 bp, Fig. 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 (Table 2, Methods M1), followed by dimensionality reduction and clustering (Fig. 1A, Methods M2). The calculated DNA structural properties included physicochemical, conformational and protein-DNA binding variables of experimental origin [5] (Table 1).

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. 2). 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 2). 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 [12, 39], horizontal flexibility [14, 15, 17, 18, 40, 41], torsional flexibility [17–19, 42], conformational stability [12, 16, 19, 20, 43], major and minor groove accessibility [18, 39, 41, 42] and A-DNA to B-DNA transition potential [17, 19, 20, 41]. Moreover, the top 20 sorted structural variables

contained 3 of the well known DNAshape functions [1, 14] and ORChID2 [13], with nucleosome positioning (phase) [44] being the most important.
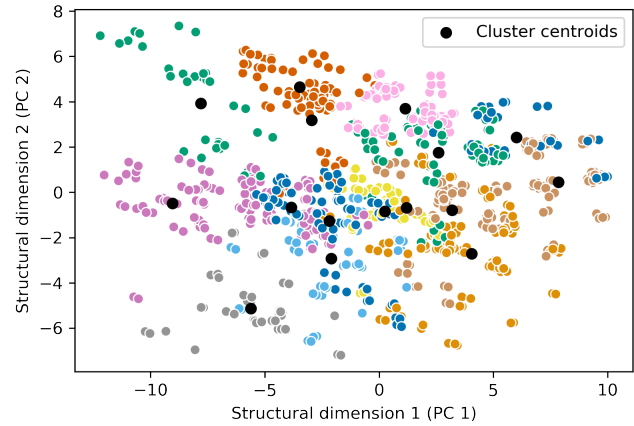

Figure 2: The first two dimensions (principal components, PC) of the structural representation with s-mer size *s* = 5 and number of clusters *k* = 16, where each point represents a different 5-mer nucleotide permutation {AAAAA, AAAAC, . . .,TTTTT}, colors depict the clusters and black points denote the cluster centroids.

Although the dimensionality reduced structural data was not expected to form strong clusters, we considered that a limit must exist to the resolution of the structural representations, above which there is no measurable influence on the achievable computational accuracy. Additionally, we wanted to achieve an additional level of compactness of the DNA codes as well as test the clustering of the structural data points. Clustering was 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 2, Methods M2). Standard cluster evaluation methods, including Elbow and Silhouette [34], 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. 3) 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. 3), with the variation of cluster sizes increasing with an increasing *k*, although the coefficient of variation did not surpass 0.52.

Table 2: Summary of s-mer construction. The groups of s-mers comprised all permutations of nucleotide k-mers of 3 to 9 bps in length. Due to sequence length constraints, not all DNA structural models could be computed with all s-mers. The amount of structural models used is given under 'Structures'. The given amount of principal components (PC) describing over 99% of the variance of the data was chosen to define each structural representation.

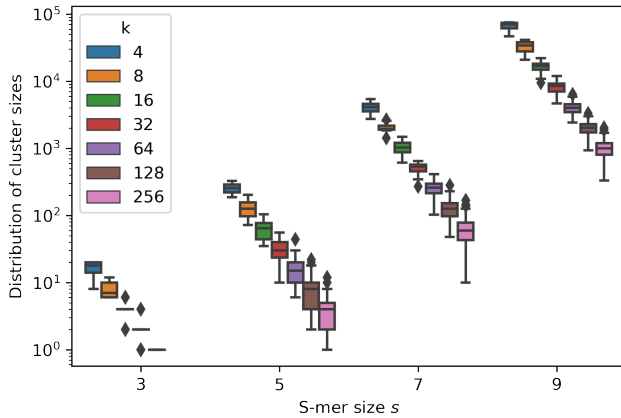| S-mer size s	Neighbors	Permutations	Structures	PC > 0.99	Number of clusters k
3	1	64	57	14	2^2 to 2^6
5	2	1024	62	17	2^2 to 2^8 + all 1024
7	3	16384	64	18	2^2 to 2^8 + all 16384
9	4	262144	64	18	2^2 to 2^8

Figure 3: The distributions of cluster sizes across the structural representations with different s-mer sizes s and number of clusters k.

3.2 Structural representations encode groups of functional sequence variants with conserved properties

We next explored whether the s-mers could encode groups of conserved functional DNA motifs [8], and if our structural DNA representations could more compactly but also more accurately encapsulate DNA motifs than bare k-mers. We used a dataset of 595 Jaspar [31] transcription factor binding site (TFBS) motifs comprising 1,296,654 unique DNA sequences from multiple model organisms (Methods M1), which contained at least 10 motifs in each group and were at least 9 bp long. To test the capacity of the structural representations to encode TFBS motifs we first measured the *compression ratio* (Eq. 3), which was the ratio of the amount of unique s-mers versus the amount of unique k-mers observed with a given TFBS motif. As a note, 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. 4), as it increased from 1.132 with $s = 5$ ($k = 256$) to 1.296 with $s = 9$ ($k = 256$). Similarly and as expected, the average compression ratio increased with decreasing amount of clusters (Fig. 4), 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.
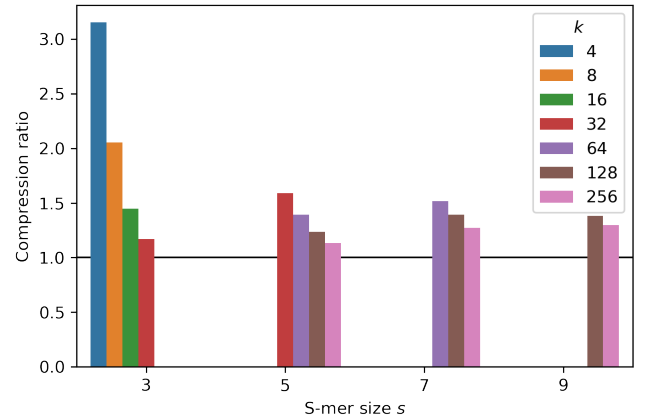

Figure 4: Compression ratios of Jaspar TFBS motifs obtained with different structural representations. Due to computation and memory limitations only certain parameter combinations of s-mer size s and number of clusters k were analysed. The black line denotes no compression.

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 M3) obtained without any encoding were 1 and 0.01, respectively. Unsurprisingly, an inverse relation was observed between *precision* and *recall* (Fig. 5). *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. 6). 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 an 18 bp DNA sequence, such as is the length of the 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. This suggested that for further analysis, datasets with a more complete coverage of the functional sequence space should be used.
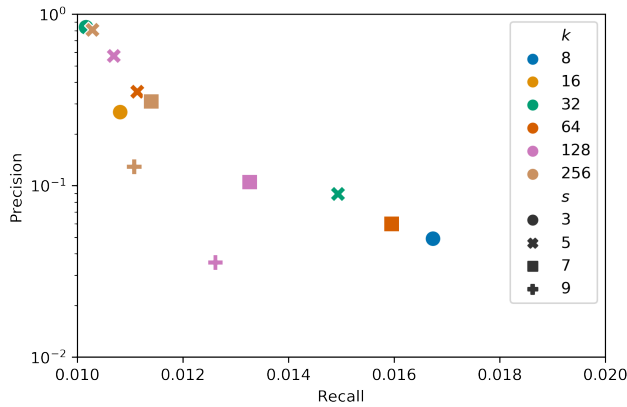

Figure 5: Precision versus recall when comparing an encoding of 1% of sequence variants of the Human MAFF motif (Jaspar: MA0495.1) with the remaining 99% of sequence variants, at different structural representation parameters s-mer size s and number of clusters k. The precision and recall obtained with nucleotide k-mers were 1 and 0.01, respectively.

3.3 A 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
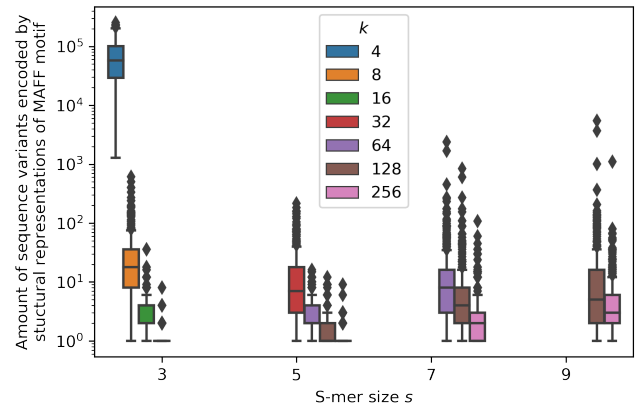

Figure 6: Amount of unique DNA sequence variants encoded by structural representations of the 18 bp long Human MAFF motif (Jaspar: MA0495.1), results with 1% (n = 100) of motif variants shown.

sequences using for instance the p-distance (equal to the Hamming distance corrected for sequence length), we defined a structural distance, termed 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, we used a dataset of whole DNA regulatory regions that control the initiation of DNA transfer in plasmid conjugation (origin-of-transfer regions) [7] (Methods 1). These transfer regions are 220 bp long and comprise binding sites for the relaxase as well as those for accessory proteins that regulate transfer. The dataset contained 64 unique transfer sequences [3] from 4 distinct mobility groups (Mob, defined by amino acid homology of the main transfer enzyme relaxase) [45], with approximately equal group sizes of 16 elements. We previously determined that the Mob groups of these regions could be clearly discriminated based on 6 DNA physicochemical and conformational properties [3]. In order to additionally test the possibility of discriminating between functional and nonfunctional transfer sequences, here we expanded the dataset to include, besides the positive sequences (Pos) also an equal amount negative counterparts (Neg, Methods M1). Using the specific distance functions as the measure of variation across the data within a permutational MANOVA framework (bootstrap n = 1e4, Methods M3) [46], we observed significant (p-value ≤ 1e-4) discrimination of both the positive/negative examples as well as MOB groups with the s-distance, which was not the case with the p-distance (Table 3). On average, when discriminating positive/negative 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 3). Additionally, using two datasets of Escherichia coli promoter regions, one with positive and negative examples [32] and the other grouped according to 6 most prevalent sigma factors [33] (Methods M1), 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

Table 3: Comparison of distance functions based on nucleotide sequence (p-distance) and structural representations (s-distance). F_1-scores were obtained with the alignment algorithm (Algorithm 1).

| Distance func.	Parameters	R^2		ANOVA p-value		F_1-score	
		Pos/Neg	Mob	Pos/Neg	Mob	Pos/Neg	Mob
p-distance	/	0.017	0.113	0.472	0.473	0.892	0.834
s-distance	$s = 3, k = 32$	0.046	0.237	< 1e-4	< 1e-4	0.955	0.853
	$s = 5, k = 128$	0.049	0.248	< 1e-4	< 1e-4	0.942	0.923
	$s = 7, k = 128$	0.054	0.253	1e-4	< 1e-4	0.970	0.923
	$s = 9, k = 128$	0.057	0.264	1e-4	< 1e-4	0.954	0.879

amount of sequence variants encoded by structural representations of the 220 bp transfer regions, by randomly selecting 10 transfer regions (Fig. 7). 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. 7).
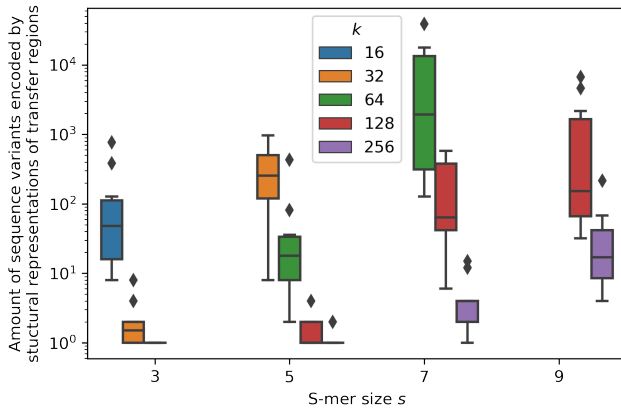
Figure 7: Amount of unique DNA sequence variants encoded by structural representations of 220 bp long plasmid transfer regions, results with 10 region variants shown.

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 DNA sequence alignment. The DNA 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 distinct transfer regions as the query dataset and a target dataset of 52 plasmids with known locations of the transfer regions as well as their Mob groups [5] (Methods M1). 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 M4) as well as the harmonic mean of precision and recall, the F_1-score (Eq. 6).

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 3). 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 [1, 8, 9, 11] (Fig. 8).
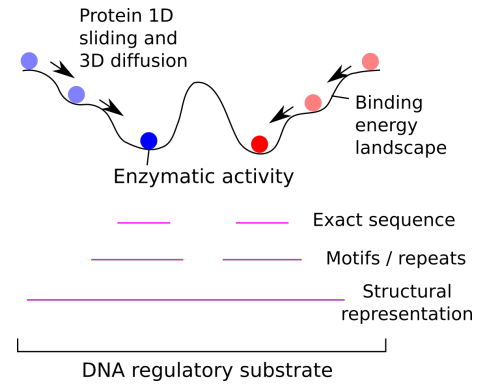
Figure 8: Possibilities for combined sequence structure algorithms based on how proteins recognize and bind their active sites in the regulatory DNA. Interactions of lower specificity with the surrounding DNA (corresponding to DNA with less conserved nucleotide sequence but defined structural properties) guide the proteins towards their specific binding sites (highly conserved or exact sequence).

4 DISCUSSION

Here we used a number of functionally-relevant DNA physico-chemical and conformational properties (Table 2: 57 to 64 structural variables) 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 1). The amount

of structural information could be further refined with clustering down to 2 bits (Fig. 3), where the compression ratio of functional DNA sequence motifs could be as high as 3:1 (Fig. 4). 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. 5). 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. 6) and up to 2000 variants of a whole DNA regulatory region (Fig. 7). 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 [3, 5] (Table 3). 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 [5], 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 (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 [5]. 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 [5].

The use of whole sets of structural variables is more frequently found in the domain of machine learning (ML), where ML models trained on specific sets of target variables then learn to use only specific subsets of the structural features [2, 3, 24]. For instance, when training ML models for discriminating Pos/Neg examples or MOB groups (Table 3), different variables at different locations were found to be most informative [3]. Although this leads to efficient discriminative or predictive models for specific tasks, DNA sequence analysis frequently requires generalizing across multiple tasks and using all of the features (e.g. alignment). Indeed, we have found that sequence based models of structural properties outperform ML models [3, 5]. Deep learning algorithms, however, might prove to be much more capable as they can interpret new data representations themselves [47] and, in our experience, outperform any structural feature-based model using mere nucleotide sequence [48].

Besides sequence alignment [5], the potential uses of the structural representations include: (i) phylogenetic analysis of regulatory DNA regions [45], (ii) analysis of single nucleotide variations [6], as the structural representations contain variants with position-specific nucleotide substitutions, (iii) motif identification

[25], where, for instance, the initialization stage of graph-based algorithms could be performed using structural representations [49], 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 [1, 8, 9, 11] and adopt a combination of both structural and sequence features (Fig. 8). Another possibility is to use different representations for the non-coding (structural) and coding (sequence) regions, e.g. for whole-gene identification. Indeed, due to the bias of the current bioinformatic methods to nucleotide sequence based approaches, it is possible that considerable performance increases might still be achievable in certain fields with such novel solutions [5, 25].

ACKNOWLEDGMENTS

REFERENCES

[1] Remo Rohs, Sean M West, Alona Sosinsky, Peng Liu, Richard S Mann, and Barry Honig. The role of DNA shape in protein–DNA recognition. *Nature*, 461(7268):1248–1253, October 2009.

[2] Jan Zrimec and Ales Lapanje. Fast prediction of DNA melting bubbles using DNA thermodynamic stability. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 12(5):1137–1145, September 2015.

[3] Jan Zrimec and Aleš Lapanje. DNA structure at the plasmid origin-of-transfer indicates its potential transfer range. *Sci. Rep.*, 8(1):1820, January 2018.

[4] Wei Chen, Pengman Feng, and Hao Lin. Prediction of replication origins by calculating DNA structural properties. *FEBS Lett.*, 586(6):934–938, March 2012.

[5] Jan Zrimec. Multiple plasmid origin-of-transfer substrates enable the spread of natural antimicrobial resistance to human pathogens. *bioRxiv*, page 2020.04.20.050401, April 2020.

[6] James D Watson, Tania A Baker, Stephen P Bell, Alexander Gann, Michael K Levine, and Richard Losick. *Molecular Biology of the Gene. 6th. ed.* Pearson/Benjamin Cummings, 2008.

[7] Fernando De La Cruz, Laura S Frost, Richard J Meyer, and Ellen L Zechner. Conjugative DNA metabolism in gram-negative bacteria. *FEMS Microbiol. Rev.*, 34(1):18–40, January 2010.

[8] Michal Levo, Einat Zalckvar, Eilon Sharon, Ana Carolina Dantas Machado, Yael Kalma, Maya Lotam-Pompan, Adina Weinberger, Zohar Yakhini, Remo Rohs, and Eran Segal. Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res.*, 25(7):1018–1029, July 2015.

[9] Amir Marcovitz and Yaakov Levy. Weak frustration regulates sliding and binding kinetics on rugged Protein–DNA landscapes. *The Journal of Physical Chemistry B*, 117(42):13005–13014, 2013.

[10] Valentina Tosato, Nicole West, Jan Zrimec, Dmitri V Nikitin, Giannino Del Sal, Roberto Marano, Michael Breitenbach, and Carlo V Bruschi. Bridge-induced translocation between NUP145 and TOP2 yeast genes models the genetic fusion between the human orthologs associated with acute myeloid leukemia. *Front. Oncol.*, 7:231, 2017.

[11] Matthew Slattery, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca Gordân, and Remo Rohs. Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.*, 39(9):381–399, September 2014.

[12] J SantaLucia, Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. U. S. A.*, 95(4):1460–1465, February 1998.

[13] Eric P Bishop, Remo Rohs, Stephen C J Parker, Sean M West, Peng Liu, Richard S Mann, Barry Honig, and Thomas D Tullius. A map of minor groove shape and

electrostatic potential from hydroxyl radical cleavage patterns of DNA. *ACS Chem. Biol.*, 6(12):1314–1320, December 2011.

[14] Tsu-Pei Chiu, Federico Comoglio, Tianyin Zhou, Lin Yang, Renato Paro, and Remo Rohs. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*, 32(8):1211–1213, April 2016.

[15] Ivan Brukner, Roberto Sanchez, Dietrich Suck, and Sandor Pongor. Sequence-dependent bending propensity of DNA as revealed by DNase i: parameters for trinucleotides. *EMBO J.*, 14(8):1812–1818, 1995.

[16] Stephanie Geggier and Alexander Vologodskii. Sequence dependence of DNA bending rigidity. *Proc. Natl. Acad. Sci. U. S. A.*, 107(35):15421–15426, August 2010.

[17] H Karas, R Knüppel, W Schulz, H Sklenar, and E Wingender. Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements. *Comput. Appl. Biosci.*, 12(5):441–446, October 1996.

[18] Wilma K Olson, Andrey A Gorin, Xiang-Jun Lu, Lynette M Hock, and Victor B Zhurkin. DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl. Acad. Sci. U. S. A.*, 95(19):11163–11168, September 1998.

[19] Alberto Perez, Agnes Noy, Filip Lankas, F Javier Luque, and Modesto Orozco. The relative flexibility of B-DNA and A-RNA duplexes: database analysis. *Nucleic Acids Res.*, 32(20):6144–6151, 2004.

[20] Misako Aida. An ab initio molecular orbital study on the sequence-dependency of DNA conformation: An evaluation of intra- and inter-strand stacking interaction energy. *Journal of Theoretical Biology*, 130(3):327–335, 1988.

[21] B Hartmann, B Malfoy, and R Lavery. Theoretical prediction of base sequence effects in DNA. experimental reactivity of Z-DNA and B-Z transition enthalpies. *J. Mol. Biol.*, 207(2):433–444, May 1989.

[22] Mandar Kulkarni and Arnab Mukherjee. Sequence dependent free energy profiles of localized B- to a-form transition of DNA in water. *J. Chem. Phys.*, 139(15):155102, October 2013.

[23] Pui S Ho, Michael J Ellison, Gary J Quigley, and Alexander Rich. A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. *EMBO J.*, 5(10):2737–2744, 1986.

[24] Jan Zrimec, Rok Kopinč, Tomaž Rijavec, Tatjana Zrimec, and Aleš Lapanje. Band smearing of PCR amplified bacterial 16S rRNA genes: dependence on initial PCR target diversity. *J. Microbiol. Methods*, 95(2):186–194, November 2013.

[25] Md Abul Hassan Samee, Benoit G Bruneau, and Katherine S Pollard. A de novo shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs. *Cell Systems*, 8(1):27–42.e6, 2019.

[26] Manju Bansal, Aditya Kumar, and Venkata Rajesh Yella. Role of DNA sequence based structural features of promoters in transcription initiation and gene expression. *Curr. Opin. Struct. Biol.*, 25:77–85, April 2014.

[27] María Lucas, Blanca González-Pérez, Matilde Cabezas, Gabriel Moncalian, Germán Rivas, and Fernando de la Cruz. Relaxase DNA binding and cleavage are two distinguishable steps in conjugative DNA processing that involve different sequence elements of the nic site. *J. Biol. Chem.*, 285(12):8918–8926, March 2010.

[28] Marta V Sut, Sanja Mihajlovic, Silvia Lang, Christian J Gruber, and Ellen L Zechner. Protein and DNA effectors control the TraI conjugative helicase of plasmid R1. *J. Bacteriol.*, 191(22):6888–6899, November 2009.

[29] Gabriel Moncalian, Mikel Valle, Jose Maria Valpuesta, and Fernando de la Cruz. IHF protein inhibits cleavage but not assembly of plasmid R388 relaxosomes. *Molecular Microbiology*, 31(6):1643–1652, 1999.

[30] Sarah L Williams and Joel F Schildbach. TraY and integration host factor orit binding sites and F conjugal transfer: sequence variations, but not altered spacing, are tolerated. *J. Bacteriol.*, 189(10):3813–3823, May 2007.

[31] Aziz Khan, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A Castro-Mondragon, Robin van der Lee, Adrien Bessy, Jeanne Chèneby, Shubhada R Kulkarni, Ge Tan, Damir Baranasic, David J Arenillas, Albin Sandelin, Klaas Vandepoele, Boris Lenhard, Benoît Ballester, Wyeth W Wasserman, François Parcy, and Anthony Mathelier. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, 46(D1):D1284, January 2018.

[32] E G Gusmão and M C P de Souto. Issues on sampling negative examples for predicting prokaryotic promoters. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 494–501. IEEE, July 2014.

[33] Socorro Gama-Castro, Heladia Salgado, Alberto Santos-Zavaleta, Daniela Ledezma-Tejeida, Luis Muñiz-Rascado, Jair Santiago García-Sotelo, Kevin Alquicira-Hernández, Irma Martínez-Flores, Lucia Pannier, Jaime Abraham Castro-Mondragón, Alejandra Medina-Rivera, Hilda Solano-Lira, César Bonavides-Martínez, Ernesto Pérez-Rueda, Shirley Alquicira-Hernández, Liliana Porrón-Sotelo, Alejandra López-Fuentes, Anastasia Hernández-Koutoucheva, Víctor Del Moral-Chávez, Fabio Rinaldi, and Julio Collado-Vides. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.*, 44(D1):D133–43, January 2016.

[34] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65, 1987.

[35] Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Series B Stat. Methodol.*, 63(2):411–423, May 2001.

[36] T D Schneider, G D Stormo, L Gold, and A Ehrenfeucht. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188(3):415–431, April 1986.

[37] Marti J Anderson. A new method for non-parametric multivariate analysis of variance. *Austral Ecol.*, 26(1):32–46, 2001.

[38] M Peyrard, S Cuesta-López, and D Angelov. Experimental and theoretical studies of sequence effects on the fluctuation and melting of short DNA molecules. *J. Phys. Condens. Matter*, 21(3):034103, January 2009.

[39] Ekaterina Protozanova, Peter Yakovchuk, and Maxim D Frank-Kamenetskii. Stacked–Unstacked equilibrium at the nick site of DNA. *Journal of Molecular Biology*, 342(3):775–785, 2004.

[40] A Bolshoy, P McNamara, and others. Curved DNA without AA: experimental estimation of all 16 DNA wedge angles. *Proceedings of the*, 6(88):2312–6, 1991.

[41] A A Gorin, V B Zhurkin, and W K Olson. B-DNA twisting correlates with base-pair morphology. *J. Mol. Biol.*, 247(1):34–48, March 1995.

[42] W Kabsch, C Sander, and E N Trifonov. The ten helical twist angles of B-DNA. *Nucleic Acids Research*, 10(3):1097–1104, 1982.

[43] M J Packer, M P Dauncey, and C A Hunter. Sequence-dependent DNA structure: dinucleotide conformational maps. *J. Mol. Biol.*, 295(1):71–83, January 2000.

[44] S C Satchwell, H R Drew, and A A Travers. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.*, 191(4):659–675, October 1986.

[45] María Pilar Garcillán-Barcia, María Victoria Francia, and Fernando de la Cruz. The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiol. Rev.*, 33(3):657–687, May 2009.

[46] Marti J Anderson. Permutational multivariate analysis of variance. *Department of Statistics, University of Auckland, Auckland*, 26:32–46, 2005.

[47] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013.

[48] J Zrimec, F Buric, A S Muhammad, R Chen, V Verendel, and others. Gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *bioRxiv*, page 10.1101/792531, 2019.

[49] Živa Stepančič. Enhancing gibbs sampling method for motif finding in DNA with initial graph representation of sequences. *Journal of Computational Biology*, 21(10):741–752, 2014.