ThreaDNA: predicting DNA mechanics' contribution to sequence selectivity of proteins along whole genomes

Jasmin Cevost, Cédric Vaillant, Sam Meyer

ThreaDNA: predicting DNA mechanics' contribution to sequence selectivity of proteins along whole genomes

Jasmin Cevost [1], Cédric Vaillant [2] and Sam Meyer [1,*]
[1] Microbiologie, Adaptation et Pathogénie, UMR5240, INSA Lyon, Université Lyon I, CNRS, Université de Lyon and
[2] Laboratoire de physique, UMR5672, ENS Lyon, CNRS, Université de Lyon

January 2018

Abstract

Motivation: Many DNA-binding proteins recognise their target sequences indirectly, by sensing DNA's response to mechanical distortion. ThreaDNA estimates this response based on high-resolution structures of the protein-DNA complex of interest. Implementing an efficient nanoscale modeling of DNA deformations involving essentially no adjustable parameters, it returns the profile of deformation energy along whole genomes, at base-pair resolution, within minutes on usual laptop/desktop computers. Our predictions can also be easily combined with estimations of direct selectivity through a generalised form of position-weight-matrices. The formalism of ThreaDNA is accessible to a wide audience.

Results: We demonstrate the importance of indirect readout for the nucleosome as well as the bacterial regulators Fis and CRP. Combined with the direct contribution provided by usual sequence motifs, it significantly improves the prediction of sequence selectivity, and allows quantifying the two distinct physical mechanisms underlying it.

Availability: Python software available at bioinfo.insa-lyon.fr, natively executable on Linux/MacOS systems with a user-friendly graphical interface. Galaxy webserver version available.

Contact: sam.meyer@insa-lyon.fr

Supplementary information: Supplementary information and data available at *Bioinformatics* online.

1 Introduction

Virtually all genomic processes are achieved by proteins which bind a subset of target DNA sequences among millions to billions of possible sites. This sequence recognition process, or readout, involves different mechanisms [1], as illustrated on Fig. 1 in the example of the bacterial architectural and regulatory protein Fis. The most natural one is a *direct* interaction of aminoacids with specific bases (G/C at positions ± 7), allowing a strong sequence selectivity. This selectivity is often described in the form of "sequence logos" (as shown on Fig. 1A) or position weight matrices (PWMs), which represent the frequency of occurrence of each nucleotide at all positions relative to the protein, as computed from a set of experimentally identified binding sites. But interestingly, in the central region of the Fis-DNA complex, depicted here in blue, the PWM also exhibits significant sequence preferences, without there being any physical contact between protein and DNA. DNA mechanics plays a central role in this *indirect* readout. Through distal interactions, Fis imposes a significant bending to DNA in this central region. Since DNA's structure and flexibility depend on its sequence, the propensity of the double-helix to accommodate the deformation will result in sequence selectivity [2]. This recognition mode is typically less specific than direct contacts, but may be particularly relevant to abundant proteins that bind many different sites along the genome, such as architectural proteins or global regulators [1]. Notably, since DNA flexibility is defined at least at the level of dinucleotides (and possibly further) rather than individual nucleotides, descriptors based on the latter, such as position weight matrices, might be poorly suited to this mechanism, as we will also show. Recent studies showed indeed that incorporating DNA shape parameters as descriptors significantly improves the prediction of regulatory protein binding sites [3]

and subsequent gene expression patterns [4].

The objective of ThreaDNA is to predict DNA mechanics' contribution to sequence selectivity of various proteins, based on a physical model of the complex without any protein-specific adjustable parameters. Such computations could be addressed with all-atomic molecular dynamics simulations [5], but these involve a considerable computational effort even for a single DNA oligomer, and cannot be used for entire genomes. Instead, ThreaDNA is based on a nanoscale description of DNA as an elastic chain of base-pairs [6]. In this computationally cheaper coarse-grained scheme, the mechanical energy associated to protein binding depends on (i) the equilibrium structure and elasticity of the DNA oligomer and (ii) its conformation in the complexed state (Fig. 1B). The former ingredient has been obtained for all possible sequences, from a combination of experimental [7] and numerical simulation [8] data. The latter ingredient is generally more difficult to obtain, because it requires some level of modeling of the protein-DNA interaction in the complex. Such models were developed mainly for the nucleosome, with various levels of complexity [9, 10, 11, 12, 13, 14, 15, 16], which demonstrated the relevance of DNA nanoscale models in predicting sequence preferences of proteins. In particular, if the protein is stiffer than DNA, it can be treated implicitly [9, 11, 13] as imposing a rigid conformation to the bound DNA, which can then be directly taken from a high-resolution of the complex, obtained by X-ray crystallography or by NMR. ThreaDNA is based on an improvement of this rigid description, which provides several key advantages as a bioinformatics tool. The very limited computational requirements of the approach make it possible to compute protein distributions on entire genomes on a desktop or laptop computer in a few minutes. In spite of this efficiency, in the case of the nucleosome, the predictive power of the algorithm matches that of more complex models developed specifically for this nucleoprotein complex (see below).

But more importantly, in contrast to previous software, ThreaDNA allows generalising the approach to a large class of DNA-binding proteins. In this paper, we illustrate the software on the important bacterial transcriptional regulators Fis and CRP. Many programs are dedicated to the prediction of transcription factor binding sites; however, most of these do not rely on physico-chemical descriptions, but rather on statistical sequence models, which must be trained on a large dataset of experimentally known binding sites, and this independently for each analysed protein. In contrast, ThreaDNA is based on validated physical models, involving essentially no free parameter, and requires no training. Its main restriction is the knowledge of a high-resolution structure of the protein-DNA complex of interest without disruption of the double-helical structure. Although this is a limiting requirement, the software remains applicable to a large class of proteins of wide interest. Importantly, to our knowledge, this is the first software to make the present approach easily accessible to the community of experimental or computational biologists interested in various proteins that distort DNA upon binding. In order to facilitate its diffusion, the algorithm was therefore implemented in a user-friendly interface designed for non-specialist users, and is also available as an installation-free online tool on the Galaxy platform [17]. The software also provides computational tools to combine the indirect contribution with direct contributions given by standard PWMs, with a strong improvement in predictive power.

2 Models and Methods

2.1 Parametrisation of DNA nanoscale elasticity

DNA's deformations are described with 6-vectors q corresponding either to base-pair or base-pair step deformations [16]. For a given sequence s, the average conformation of naked DNA is quoted $q_0(s)$, and its rigidity is given by the stiffness matrix $K(s)$. The software includes two parameter sets ("NP" or "ABC"), and the length of s (di/tri/tetranucleotide) depends on this choice. The "NP" parameter set describes the elasticity of base-pair steps at the dinucleotide level, and was obtained from a collection of high-resolution DNA or DNA-protein crystallographic structures [7]. In the "ABC" parameter set, obtained from microsecond-long molecular dynamics simulations [8], the base-pair step elasticity depends on the tetranucleotide sequence ("ABC_s"), and the internal base-pair deformations on the trinucleotide sequence ("ABC_i"). In all cases, under the elastic approximation, the associated deformation energy follows the same quadratic form: $U(q,s) = \frac{1}{2}(q - q_0(s))^t K(s)(q - q_0(s))$.

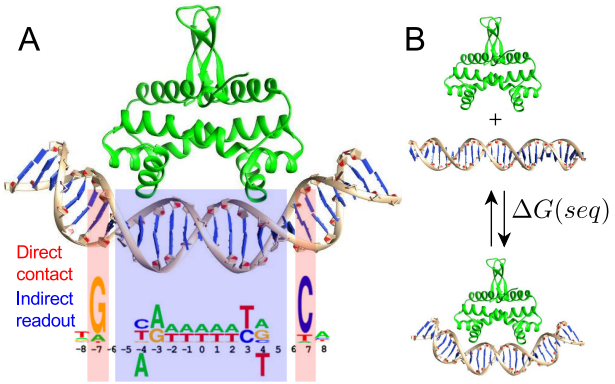

Figure 1: Sequence recognition mechanisms illustrated on the bacterial nucleoid-associated protein Fis. **(A)** Fis directly contacts the DNA bases at positions ± 7 (red background), where it exhibits a strong preference for G/C bases. In the central region (blue), DNA is not contacted, but Fis still exhibits significant sequence preferences, possibly for sequences most favourable to Fis-induced bending. Adapted from [18]. **(B)** ThreaDNA estimates DNA mechanical energy associated to protein binding from high-resolution structures of the complex, which provides a significant contribution to sequence selectivity.

2.2 Estimation of sequence-dependent deformation energy from protein-DNA structural models

The deformation of a given DNA base-pair (step) conformation is the result of a mechanical equilibrium between the external potential exerted by the protein $F_{ext}(q)$, and the DNA internal elastic energy $U(q, s)$. Even in the case of non-specific binding ($F_{ext}(q)$ independent of the sequence), the resulting equilibrium conformation q_e minimising the total energy still depends on the sequence, and so does also the deformation energy, see details in [16]. Computing these conformations generally requires an involved modeling of this external force field [10, 15, 16]. To simplify the calculation, we take advantage of the strong stiffness of many DNA-binding proteins, as compared to DNA stiffness. In the limit of infinite stiffness, the conformation q_e becomes sequence-independent, and the deformation free energy can be computed very efficiently from the formula above. This approach (hereafter referred to as a "rigid approach") has been proposed for the nucleosome [9, 11, 7, 13]. Its main limitation is that the histone core is actually not infinitely stiff: the observed structures depend quite strongly on the incorporated sequence [19], and the predictions of these programs depend equally strongly on which structural model is used as template (see Results). Our algorithm overcomes this limitation by *combining* the whole dataset of known structures into a single computation of deformation energy, as follows. If different conformations $\{q_e^i\}_{i=1,\dots,P}$ were observed in different crystals, the calculation is based on the hypothesis that this set is representative of the configurational space accessible to the complex incorporating any sequence. For a given sequence s, the equilibrium conformation q adopted will then be a statistical "mix" of these states, with a respective weight given by the Maxwell-Boltzmann equilibrium statistics: the total deformation free energy $F(s)$ is then given by $\exp(-\beta F(s)) = \sum_{i=1}^{P} \exp(-\beta U(q_e^i, s))$, where $\beta = 1/(k_B T)$ is an effective temperature parameter. Our algorithm extends the method to other proteins, where direct as well as indirect readout contribute to the binding free energy.

2.3 Implementation of algorithm

The algorithm was written in Python2 using the NumPy library [20] and the Tkinter cross-platform graphical library. This makes the code easy to change, and natively executable on recent Linux or MacOS systems, as well as on Windows after installing these tools. The user-friendly graphical interface is designed for non-specialist users focusing on a protein of interest; and the program is numerically optimised so as to allow genome-wide computations on a desktop computer. The first (and delicate) step is the inclusion of the protein-DNA complex of interest into the software database. It can be provided either in the form of a NDB ID [21], or an output file of the popular conformational analysis programs 3DNA [22] (available on the webserver Web3DNA) or Curves+ [23]. For new proteins, it is crucial to check that the DNA structure was not broken during the conformational analysis, otherwise the computation will be meaningless; usually, one only has to check that no base-pair is missing in the base-pair step list. A previously observed difficulty in the analysis is that, because of thermal and experimental noise in the structures, the computed energy profiles are affected by an unknown scaling factor ("effective temperature") [24, 16]: combining different structures of different non-thermal energy scales thus imposes an arbitrary rescaling choice. We choose the normalising constant such that, for a random sequence, the profiles computed with all structural models have the same standard deviation. An "effective

temperature" parameter then fixes the global energy scale: for the default value of 1, all incorporated structural models contribute significantly to the combined profile, and a smaller value gives more weight to the most favourable structures. As a result, the computed profiles are all given in arbitrary units. Details of the computation parameters are described in the software documentation. The memory requirement and computation time is proportional to the protein size (in contacted base-pairs) and genome length.

3 Results

3.1 ThreaDNA: a new tool to analyse physical features involved in protein binding at the genomic scale

ThreaDNA is a new software that predicts the contribution of DNA deformations in the sequence selectivity of proteins. Such deformations are often described extensively in structural studies [27, 18], but their translation into sequence-dependent binding preferences requires a suitable physical model of the DNA-protein complex. The efficient "rigid" approach proposed by ThreaDNA relies on a nanoscale description whereby the protein imposes the conformation of the DNA base-pairs in the complex [9, 11, 13]. This approximation considerably simplifies the analysis of the system, which can then be implemented by a non-specialist user for a large class of DNA-binding proteins of interest, provided high-resolution structures of the complex involving no double-helix disruption are available from X-ray crystallography or NMR experiments. The inclusion of several alternate structural models for the same protein results in a refined physical description of the interaction, while preserving its strong computational efficiency. As an example, for the proteins described in this paper, computing DNA deformation energies at all possible positions along a bacterial or yeast genome of a few megabases typically takes less than a minute on a laptop or desktop computer, using less than 1 Gb RAM.

In a first step, ThreaDNA requires the user to incorporate the DNA conformation of the nucleoprotein complex of interest into the software database. We recommend using the Web3DNA webserver [22] to analyse a PDB file of interest. The output (.out) file can be provided to ThreaDNA together with the original PDB file. In a second step, the main program can then use this conformation to compute mechanical energies on any sequence provided in the standard Fasta format. The details of the calculation (list of considered structures, DNA stiffness parameter set...) are specified in an input file of standard format.

The program can be executed via a user-friendly graphical interface, available as a Supplementary File, which runs natively on MacOS/Linux distributions. Additionally, a web-server version is available on a local version of the Galaxy bioinformatics platform (`bioinfo.insa-lyon.fr`) [17], which provides an installation-free access to all users. The returned profiles are in the standard BedGraph format and can be visualised on genome browsers (Fig. S1). The indirect selectivity of the protein is also returned as a generalised position weight matrix (PWM), that represents the energy or occurrence probability for each di/tetranucleotide along the protein (see below), in the standard JASPAR format. An additional subprogram facilitates the manipulation and combination of such generalised PWMs.

3.2 The nucleosome: genomic and high-affinity sequences

To illustrate the predictive power and possible applications of the software, we start with the nucleosome. Since most previous comparable models were developed specifically for this complex, we use it as a benchmark, keeping in mind that our software aims at generalising the approach to other proteins. We first use ThreaDNA to compute nucleosome occupancies on the entire *Saccharomyces cerevisiae* genome (Fig. 2A). These predictions are compared to experimental data obtained from DNA digestion experiments by micrococcal nuclease followed by high-throughput sequencing [25], either on *in vitro* reconstituted chromatin, or on the native ("*in vivo*") fibre. The calculation involves computing (i) deformation energies at all genomic positions without any free parameter, and (ii) a Boltzmann inversion of the profile to get occupancies, using a single global scaling parameter (effective temperature). Note that the latter operation coarsely neglects the interaction between adjacent nucleosomes [28, 29]; this is justified here by our focusing on the former computation by the present software. The agreement is good with *in vitro* data ($r^2 = 0.56$), and only slightly less so with *in vivo* data, suggesting that in yeast, the positions occupied by most nucleosomes are thermodynamically encoded in the sequence through DNA mechanical properties. These results are comparable to previous ones obtained by rigid approaches, *e.g.*, a correlation of 0.45 is reported in [30]. In contrast, better correlations could only

be obtained using a different class of sequence-based models trained on the analysed data [25, 29], involving a large number of fitted parameters.

However, it was noted that, with some exceptions [29], even these specialised models fail to reproduce the strong affinity of specific sequences used in crystallisation experiments. A possible explanation is that these sequences bind the histones with qualitatively different conformations [19], which is difficult to take into account within a single model. Notably, ThreaDNA takes this plasticity into account, by letting each sequence statistically "choose" its favourite shape from a combination of crystallised structures, rather than imposing a single one. Although this treatment is simplified compared to previous schemes [10, 15, 16], it successfully reproduces the strong affinity of various sequences (Fig. 2B, lower panel), which is not possible based on a single structural template (upper panel and Fig. S3) [11, 12, 13]. Although the structural landscape sampled by ThreaDNA may not be comprehensive with respect to other high-affinity or genomic sequences of yet unknown conformation [16], new structures will likely be obtained regularly in the future, thereby "automatically" refining our model. Altogether, considering the simplifications imposed by its purpose, our model stands the comparison with most reported predictive models dedicated specifically to the nucleosome, including for the delicate prediction of high-affinity sequences. But more importantly, our program is the first to generalise the approach to other proteins, for which none is currently available.

3.3 Sequence selectivity of Fis, a major bacterial architectural and regulatory protein

Factor for inversion stimulation (Fis) is one of the most abundant DNA-binding proteins in *Escherichia coli* in conditions of rapid growth [32]. As shown in Fig. 1, Fis binds around 15 bp of DNA, with specific contacts at positions ± 7, and indirect sequence readout occurring in the central region. It has been suggested that the latter recognition might be achieved through DNA denaturation [33]; however, since existing crystallographic models of the complex exhibit significant DNA bending without disruption of the double-helix, we address the question, whether double-stranded flexibility of DNA might rather be the relevant mechanism. To answer this question, we used ThreaDNA to compute deformation energies imposed by Fis at all positions along the *E. coli* chromosome, and asked if experimentally known binding sites are associated to significantly low values.

Fis binding sites were detected in many separate studies, each of them often focusing on a particular gene promoter, and this information was collected into the RegulonDB database [26]. The database thus contains around 250 binding sites at base-pair resolution, with different levels of confidence depending on the experimental method used. Fig. 3 shows that the average mechanical energy profile around experimental binding sites (blue) drops at the binding position, indicating that Fis binding sequences are associated to weaker mechanical energies than their neighbours, and than genomic average. Consistently, the associated histogram is significantly shifted to the left (p-value $P < 10^{-6}$, Kolmogorov-Smirnov test), as compared to the whole genome. This observation is expected if DNA mechanical properties play a significant role in the selectivity, as hypothesised. On the other hand, we also note that the deformation energies of most experimental binding sites are not among the lowest in the genome, indicating that other mechanisms are involved in the recognition process, including the aforementioned direct readout at positions ± 7.

Fis was crystallised in complex with several DNA sequences, which adopt different conformations [18]. We analysed this flexibility by computing the deformation energy associated to these sequences, depending on which conformation is used as template in the program (Fig. S4). In most cases, ThreaDNA correctly predicts the observed binding position, associated to a low deformation energy, when the computation is based on the specific conformation adopted by the considered sequence, but not necessarily by other conformations. Thus, as observed previously for the nucleosome, the conformational freedom exhibited in the crystals is an important ingredient for the prediction of high-affinity sequences. Indeed, the prediction of binding affinities for these sequences is strongly improved by the use of a combined structure ($r^2 = 0.44$) compared to a single structure ($r^2 = 0.3$).

3.4 DNA sequence-encoded mechanical properties are essential determinants of CRP selectivity

cAMP receptor protein (CRP) is another important global bacterial regulator, involved in cell metabolism. It binds several hundred sites along the *E. coli* genome, also with a combination of direct and indirect sequence recognition [27], and bends the DNA by al-

most 90° (Fig. 4A). The degeneracy of recognised sequences in the non-contacted central region thus also points to a role for DNA mechanics. However, in absence of physical models of the interaction, little is known on this contribution.

We start with the approach already used for Fis, with the ∼350 experimental binding sites listed in the RegulonDB database, and the 1CGP structural model for DNA deformations [35]. The results are shown on Fig. 4: the decrease observed at experimental positions is here much stronger than for Fis ($P < 10^{-16}$, Kolmogorov-Smirnov test). Importantly, in contrast to the latter case, the large majority of observed values are among the weakest of the genome. This observation suggests that DNA distortions could not only be a significant, but even the dominant contribution to the sequence selectivity of CRP. Yet this hypothesis seems contradicted by the presence of a minority of datapoints in the right side of the histogram. We reasoned that these points might also correspond to false positives of the database, as RegulonDB incorporates data of heterogeneous levels of confidence, and considering that the spatial resolution of most methods used (including DNA footprinting by DNAse I) is lower than a single basepair. As an example, two gene regulatory regions (*sohB* and *proP*) reportedly contain two binding sites separated by a single basepair, which seems unlikely. In both cases, only one of the two sites has a low bending energy, whereas the other is very unfavourable (identifiers ECK120012440 and ECK120015998), likely indicative of false positives. The same is true of 6 out of 11 reported regions with overlapping CRP sites.

To weaken the effect of such issues, we repeated the analysis on an independent dataset obtained by genomic SELEX [31]. In contrast to the previous case, these sites are obtained from a single genome-wide experiment, where DNA fragments enriched in CRP binding are isolated, and a subsequent bioinformatics analysis allows identifying the strongest binding sites at base-pair resolution. The resulting list is shorter than that of RegulonDB (∼260 sites), but since it does not result from a mixing of heterogeneous studies, it might include less false positives. Indeed, the histogram is quite similar (Fig. 4C), but remarkably, the group of datapoints on the right part has almost entirely vanished. Taken together, these observations clearly show that CRP binding sites are systematically characterised by very low deformation energies, suggesting an important or even dominant role for indirect readout.

We now quantitatively test the predictive power of the software on a group of 25 mutants of the high-affinity LacP1 sequence from *E. coli*, whose affinities for CRP were measured *in vitro* by [34]. The mutants were selected either randomly, or for their high affinity (with a maximal 14-fold change). Importantly, mutations all occurred in the central, non-contacted region of the DNA oligomer. It is thus reasonable to assume that direct readout contributes only marginally to these affinity variations, which can then be directly compared to the predictions of ThreaDNA. For all sequences, we found the lowest deformation energy at the experimental binding position (data not shown). Fig. 5A shows that these deformation energies correlate well with the measured affinities (correlation coefficient $r^2 = 0.53$, p-value $P = 2.6 \times 10^{-5}$). Remarkably, this correlation coefficient is much higher than that obtained with the position weight matrix (PWM) constructed with all RegulonDB sites ($r^2 = 0.3$, see also Fig. S6). Thus, even though our model is based on structural information only and involves no training on CRP binding sequences, in this indirect readout assay, it outperforms a PWM that concentrates our knowledge of CRP binding selectivity obtained through hundreds of studies. Note that although the PWM theoretically provides binding affinities in absolute units, in practice we had to rescale them as we did for ThreaDNA, using a single slope parameter.

At this point, the reader should keep in mind that since ThreaDNA tells nothing on direct recognition, its general objective is not to compete with sequence motifs, but rather to pinpoint and quantify the indirect contribution to the binding selectivity represented in these motifs. In the next paragraph, we show that this information can be combined with usual PWMs containing complementary information about direct readout, with a notable improvement in predictive power.

3.5 Combining indirect and direct readout significantly improves the prediction of binding selectivity

Our previous results immediately suggest to combine our estimation of the indirect contribution with the PWM describing the total (direct+indirect) interaction, in order to infer the direct contribution. Theoretically, this can be achieved by computing a new PWM, where each nucleotide is weighted according to its indirect contribution. In practice, this "subtraction" operation is challenging: for instance, any wrong binding site in the list (false positive), associated to a high deforma-

tion energy, will get an enormous weight in the new PWM. Fortunately, it is possible to avoid this difficulty, by noticing that the indirect contribution computed from ThreaDNA acts at least at the level of dinucleotides, as represented in Fig. 5C (lower panel), rather than mononucleotides. The software provides such generalised "di-PWMs", which can be manipulated and combined with usual PWMs with a subprogram. Importantly, if we try to project this di-PWM into a traditional mono-PWM, we lose most of its predictive power ($r^2 = 0.2$), which shows that indirect readout *cannot* be described at the mononucleotide level. As a consequence, it is a reasonable approximation to consider the usual mono-PWM as a representation of the direct contribution rather than the total interaction. In other words, the ThreaDNA di-PWM and the traditional mono-PWM can be considered as providing orthogonal information about indirect and direct readout contributions respectively, which can then simply be added to compute the total interaction. Their relative weight cannot be predicted, but may be estimated by comparison to the data. For the Lindemose et *al*. sequences, the combination remarkably improves the predictive power, with up to 0.88 correlation with the experimental values for 35% direct readout ($P < 10^{-11}$, Fig. 5B). To confirm that this result is not an accident, we repeated the analysis on Fis, where the combination also improves the results significantly (Fig. S5). The lower correlation values compared to CRP might be due to the mutations occuring in the regions in direct contact with the protein, which may be more difficult to describe.

4 Discussion

In order to recognise their target sequences, DNA-binding proteins often use a combination [1] of direct readout of DNA bases and indirect readout of the "analogue" mechanical code encoded in the DNA sequence [36]. ThreaDNA implements a systematic approach to estimate the latter contribution for a large class of nucleoprotein complexes. Being based on a nanoscale physical analysis of structural data, it does not require any training on protein-specific binding datasets. All essential model parameters are protein-independent and chosen from known experimental values. Starting from a structural model of the protein-DNA complex, the software provides genome-wide deformation energy profiles within minutes on standard computers.

We wish to give some emphasis on the assumptions and limitations of ThreaDNA, in order to avoid erroneous applications of the software. Firstly, the only DNA deformations analysed are those located within the structural models provided as input. As a consequence, regarding proteins that bind together distant DNA sites or act collectively, *e.g.* by bridging, looping, polymerisation, etc., ThreaDNA might give useful information on their individual binding affinity, but tells nothing of the collective process, for which additional detailed modeling of the interaction is required [37, 28]. The main assumption of the model is that the elasticity of (naked) double-helical DNA remains valid within the protein-DNA complexes analysed, which is incompatible with proteins that disrupt the Watson-Crick bonds of the base-pairs or distort them too strongly. Such problems can generally be identified in the form of missing or aberrant values in the list of base-pair step geometrical parameters given in the Nucleic Acids Database [21]. As noted earlier, caution must therefore be paid upon including a new structure into the database.

Still, previous analyses suggest that the software can be useful even for proteins exerting extreme constraints on DNA. One such example is the HIV integrase, a large nucleoprotein complex that inserts viral DNA into the genome of a host human cell, with non-uniform insertion preferences. Using a structural model of the pre-integration complex [38] exhibiting significant deformations of the genomic DNA prior to its disruption, we predicted insertion preferences in surprisingly good agreement with the inhomogeneous patterns observed both *in vitro* and *in vivo* [39]. A possible interpretation is that indirect readout might be used in intermediate kinetic steps by proteins that end up disrupting or even breaking the double-helix. This seems a reasonable hypothesis in the speculative point of view of a protein which has to find its targets among millions of possible sites: the "cheap" (low energy-barrier) DNA deformation steps might contribute in "preselecting" favourable sequences, while "expensive" base-pair disruption occurs only as a less frequent second step. Although these kinetic views go beyond our thermodynamic approach, a better understanding of the latter might prove helpful in deciphering the whole process.

Our analysis shows that ThreaDNA may substantially improve the prediction of binding sequences for proteins that distort DNA, by complementing the usual approach based on position weight-matrices. The software includes Python modules that facilitate their combination. It is well-known that many proteins use a combination of direct and indirect readout [1], but

to our knowledge, our approach is the first to explicitly use this distinction for quantitative sequence binding predictions. In addition to practical applications in binding selectivity prediction for various proteins, it will therefore also provide a new understanding of the mechanistic features underlying this selectivity.

5 FUNDING

This work was supported by a Bonus de Qualité Recherche (BQR) INSA Lyon 2016 grant, and an IXXI 2016 grant.

6 ACKNOWLEDGEMENTS

We thank Caroline Gaud, Adelme Bazin and Adeline Roatta who contributed to the implementation and testing of the software, Patrice Baa-Puyoulet for his involvement, Ralf Everaers for useful discussions, William Nasser for his support and advice, and Sylvie Reverchon and Georgi Muskhelishvili for their critical reading of the manuscript.

6.0.1 Conflict of interest statement.

None declared.

References

[1] Sabrina Harteis and Sabine Schneider. Making the bend: DNA tertiary structure and protein-DNA interactions. *Int J Mol Sc*, 15(7):12335–12363, 2014.

[2] Nils B Becker, Lars Wolff, and Ralf Everaers. Indirect readout: detection of optimized subsequences and calculation of relative binding affinities using different DNA elastic potentials. *NAR*, 34(19):5638–5649, 2006.

[3] Tianyin Zhou, Ning Shen, Lin Yang, Namiko Abe, John Horton, Richard S Mann, Harmen J Bussemaker, Raluca Gordân, and Remo Rohs. Quantitative modeling of transcription factor binding specificities using DNA shape. *PNAS*, 112(15):4654–4659, 2015.

[4] Pei-Chen Peng and Saurabh Sinha. Quantitative modeling of gene expression using DNA shape features of binding sites. *NAR*, 44(13):e120, 2016.

[5] Guillaume Paillard and Richard Lavery. Analyzing protein-DNA recognition mechanisms. *Structure*, 12(1):113–122, 2004.

[6] Sam Meyer, Daniel Jost, Nikos Theodorakopoulos, Michel Peyrard, Richard Lavery, and Ralf Everaers. Temperature Dependence of the DNA Double Helix at the Nanoscale: Structure, Elasticity, and Fluctuations. *BPJ*, 105(8):1904–1914, 2013.

[7] Fei Xu and Wilma K. Olson. DNA architecture, deformability, and nucleosome positioning. *J Biomol Struct Dyn*, 27(6):725–739, June 2010.

[8] Marco Pasi, John H. Maddocks, David Beveridge, Thomas C. Bishop, David A. Case, Thomas Cheatham, Pablo D. Dans, B. Jayaram, Filip Lankas, Charles Laughton, Jonathan Mitchell, and others. muABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *NAR*, 42(19):12272–12283, 2014.

[9] C. Anselmi, G. Bocchinfuso, P. De Santis, M. Savino, and A. Scipioni. A theoretical model for the prediction of sequence-dependent nucleosome thermodynamic stability. *BPJ*, 79(2):601–613, 2000.

[10] Alexandre V Morozov, Karissa Fortney, Daria A Gaykalova, Vasily M Studitsky, Jonathan Widom, and Eric D Siggia. Using DNA mechanics to predict in vitro nucleosome positions and formation energies. *NAR*, 37(14):4707–22, August 2009.

[11] Fei Xu, Andrew V. Colasanti, Yun Li, and Wilma K. Olson. Long-range effects of histone point mutations on DNA remodeling revealed from computational analyses of SIN-mutant nucleosome structures. *NAR*, 38(20):6872–6882, November 2010.

[12] Richard C. Stolz and Thomas C. Bishop. ICM Web: the interactive chromatin modeling web server. *Nucleic Acids Research*, 38(suppl_2):W254–W261, July 2010.

[13] Ö. Deniz, O. Flores, F. Battistini, A. Pérez, M. Soler-López, and M. Orozco. Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast. *BMC genomics*, 12(1):489, 2011.

[14] Pasquale De Santis and Anita Scipioni. Sequence-dependent collective properties of

DNAs and their role in biological systems. *Phys Life Rev*, 10(1):41–67, 2013.

[15] Arman Fathizadeh, Azim Berdy Besya, Mohammad Reza Ejtehadi, and Helmut Schiessel. Rigid-body molecular dynamics of DNA inside a nucleosome. *Eur. Phys. J. E*, 36(3):1–10, 2013.

[16] Sam Meyer and Ralf Everaers. Inferring coarse-grain histone-DNA interaction potentials from high-resolution structures of the nucleosome. *J Phys: Cond Mat*, 27(6):064101, 2015.

[17] Enis Afgan, Dannon Baker, Marius van den Beek, Daniel Blankenberg, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Carl Eberhard, and others. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *NAR*, 44(W1):W3–W10, 2016.

[18] Stefano Stella, Duilio Cascio, and Reid C Johnson. The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. *Genes & development*, 24(8):814–826, 2010.

[19] W.K. Olson and V.B. Zhurkin. Working the kinks out of nucleosomal DNA. *Curr Op Struct Biol*, 2011.

[20] Stefan Van Der Walt, S Chris Colbert, and Gael Varoquaux. The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2):22–30, 2011.

[21] Buvaneswari Coimbatore Narayanan, John Westbrook, Saheli Ghosh, Anton I Petrov, Blake Sweeney, Craig L Zirbel, Neocles B Leontis, and Helen M Berman. The Nucleic Acid Database: new features and capabilities. *NAR*, 42(D1):D114–D122, 2014.

[22] Xiang-Jun Lu and Wilma K Olson. 3dna: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *NAR*, 31(17):5108–5121, September 2003.

[23] R. Lavery, M. Moakher, JH Maddocks, D. Petkeviciute, and K. Zakrzewska. Conformational analysis of nucleic acids revisited: Curves+. *NAR*, 37(17):5917–5929, 2009.

[24] Nils B Becker and Ralf Everaers. DNA nanomechanics in the nucleosome. *Structure*, 17(4):579–589, April 2009.

[25] Noam Kaplan, Irene K Moore, Yvonne Fondufe-Mittendorf, Andrea J Gossett, Desiree Tillo, Yair Field, Emily M LeProust, Timothy R Hughes, Jason D Lieb, Jonathan Widom, and Eran Segal. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236):362–366, March 2009.

[26] 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. *NAR*, 44(D1):D133–D143, 2016.

[27] Andrew A Napoli, Catherine L Lawson, Richard H Ebright, and Helen M Berman. Indirect Readout of DNA Sequence at the Primary-kink Site in the CAP–DNA Complex: Recognition of Pyrimidine-Purine and Purine-Purine Steps. *Journal of molecular biology*, 357(1):173–183, 2006.

[28] G Chevereau, L Palmeira, C Thermes, A Arneodo, and C Vaillant. Thermodynamics of intragenic nucleosome ordering. *PRL*, 103(18):188103, October 2009.

[29] Thijn van der Heijden, Joke J F A van Vugt, Colin Logie, and John van Noort. Sequence-based prediction of single nucleosome positioning and genome-wide nucleosome occupancy. *PNAS*, 109(38):E2514–22, September 2012.

[30] Vincent Miele, Cedric Vaillant, Yves d'Aubenton Carafa, Claude Thermes, and Thierry Grange. DNA physical properties determine nucleosome occupancy from yeast to fly. *NAR*, 36(11):3746–3756, 2008.

[31] Tomohiro Shimada, Nobuyuki Fujita, Kaneyoshi Yamamoto, and Akira Ishihama. Novel roles of cAMP receptor protein (CRP) in regulation of transport and metabolism of carbon sources. *PLoS One*, 6(6):e20081, 2011.

[32] Andrew Travers, Robert Schneider, and Georgi Muskhelishvili. DNA supercoiling and transcription in Escherichia coli: The FIS connection. *Biochimie*, 83(2):213–217, 2001.

[33] Kristy Nowak-Lovato, Ludmil B Alexandrov, Afsheen Banisadr, Amy L Bauer, Alan R Bishop, Anny Usheva, Fangping Mu, Elizabeth Hong-Geller, Kim Ø Rasmussen, William S Hlavacek, and others. Binding of nucleoid-associated protein Fis to DNA is regulated by DNA breathing dynamics. *PLoS Comput Biol*, 9(1):e1002881, 2013.

[34] Søren Lindemose, Peter Eigil Nielsen, Poul Valentin-Hansen, and Niels Erik Møllegaard. A novel indirect sequence readout component in the E. coli cyclic AMP receptor protein operator. *ACS Chem Biol*, 9(3):752–760, 2014.

[35] SC Schultz, GC Shields, and TA Steitz. Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science*, 253(5023):1001–1007, 1991.

[36] AA Travers, G. Muskhelishvili, JMT Thompson, AA Travers, G. Muskhelishvili, and JMT Thompson. DNA information: from digital code to analogue structure. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1969):2960–2986, 2012.

[37] Vladimir B. Teif. General transfer matrix formalism to calculate DNA–protein–drug binding in gene regulation: application to OR operator of phage lambda. *Nucleic Acids Research*, 35(11):e80, 2007.

[38] Fabrice Michel, Corinne Crucifix, Florence Granger, Sylvia Eiler, Jean-François Mouscadet, Sergei Korolev, Julia Agapkina, Rustam Ziganshin, Marina Gottikh, Alexis Nazabal, and others. Structural basis for HIV-1 DNA integration in the human genome, role of the LEDGF/P75 cofactor. *The EMBO journal*, 28(7):980–991, 2009.

[39] Monica Naughtin, Zofia Haftek-Terreau, Johan Xavier, Sam Meyer, Maud Silvain, Yan Jaszczyszyn, Nicolas Levy, Vincent Miele, Mohamed Salah Benleulmi, Marc Ruff, and others. DNA Physical Properties and Nucleosome Positions Are Major Determinants of HIV-1 Integrase Selectivity. *PloS one*, 10(6), 2015.

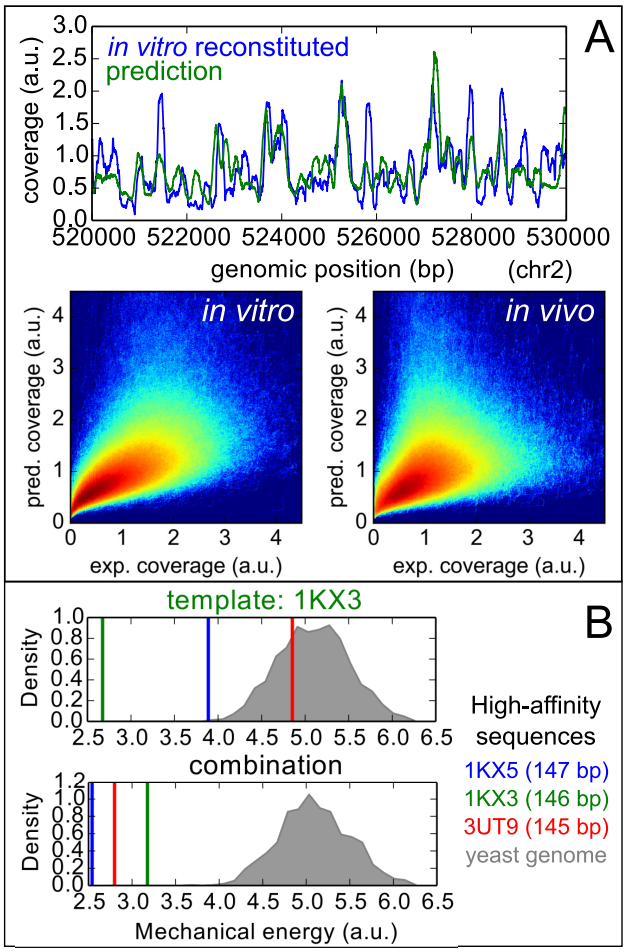

Figure 2: ThreaDNA provides efficient predictions of nucleosome binding preferences encoded in DNA elasticity. **(A)** Comparison of nucleosome occupancy profiles predicted by ThreaDNA along the *S. cerevisiae* genome with MNAse-seq profiling of nucleosomes on *in vitro* reconstituted or *in vivo* (native) chromatin [25]. Upper panel: illustration on a particular location of chromosome 2. Lower panels: Correlation histograms along the whole genome (logarithmic density scale): linear correlation coefficients 0.56 and 0.36 respectively, both p-values $P < 10^{-16}$. **(B)** Prediction of wrapping energies of high-affinity sequences. While a single structural template (here 1KX3) overestimates the energy of other high-affinity sequences, ThreaDNA combines their structures into a single model (lower panel), which accurately predicts low energies for all sequences. All calculations involved the NP parameter set (see Figs. S2 and S3 for comparison).

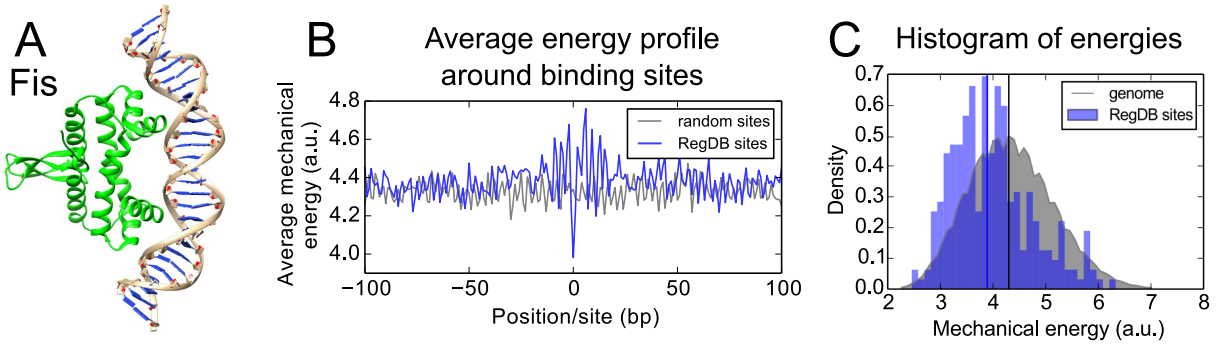

Figure 3: DNA mechanics provides a significant contribution to the sequence selectivity of Fis in *Escherichia coli*. **(A)** Crystallographic structural model 3IV5 of the Fis-DNA complex [18]. **(B)** Average mechanical energy profile around experimental binding sites collected in the RegulonDB database [26] (blue) vs random sites (grey). Obtained with 3JR9 structure [18] and ABC step parameter. **(C)** The histogram of mechanical energies at experimental binding sites (position 0 in B) is significantly shifted toward low energies ($P < 10^{-6}$). Still, a large fraction of these sites exhibits medium or high energies, confirming that other mechanisms are involved in Fis sequence selectivity (Fig. 1).

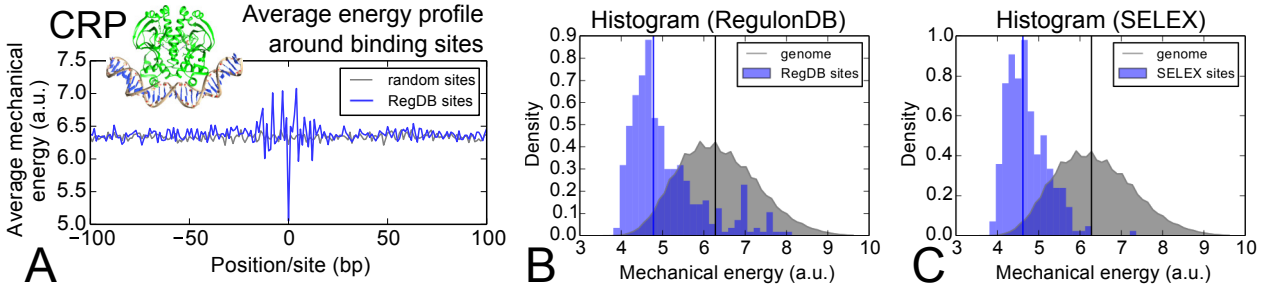

Figure 4: Role of indirect readout in CRP sequence recognition. **(A)** The deformation energy is strongly lower at experimental CRP binding sites (from RegulonDB), compared to random sites (same legend as Fig. 3). **(B)** Associated histogram of energies (position 0 in A): in contrast to Fis, DNA deformation energy is very low at almost all CRP binding sites. **(C)** Binding sites identified by SELEX [31], possibly incorporating less false positives, the group of sequences with high deformation energies has almost vanished.

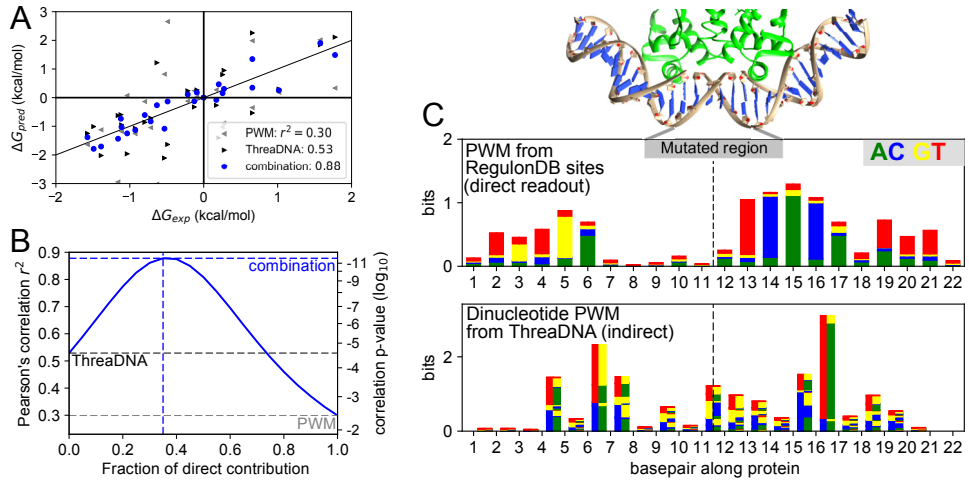

Figure 5: (A) Comparison of *in vitro* experimental affinities for mutants of the high-affinity LacP1 sequence [34] with predictions of the position weight matrix (PWM) based on all RegulonDB sites, of ThreaDNA, and of a combination of the two. The reference energy is for the wild-type LacP1 site. The arbitrary unit of ThreaDNA results was converted into kcal/mol using the slope of the regression (one adjusted parameter), and the PWM predictions were equally rescaled (by a factor of 2.5). The computation was carried with a combination of the 1CGP and 1ZRF structures [35, 27], with the NP parameter set. (B) Combining the indirect (ThreaDNA) and direct (PWM) contributions to sequence selectivity considerably improves the prediction. In this assay, a fraction of $\sim 35\%$ direct readout contribution gives the best results (shown in A). (C) In our combination scheme, the direct contribution is approximated by the "traditional" PWM (see text) obtained from all RegulonDB binding sites of CRP (top panel), while the indirect contribution is given by a generalised "dinucleotide" step PWM provided by ThreaDNA. The grey box indicates the mutated region in the Lindemose et *al*. study.