COMMENTARY

# Shapely DNA attracts the right partner

**Teresa M. Przytycka[a,1] and David Levens[b,1]**

[a]*Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20892; and* [b]*Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892*

All levels of cell activity are coordinated directly or indirectly by transcription factors (TFs). In turn, the functioning of TFs relies on their ability to recognize and bind specific DNA sequences to regulate the expression of specific genes. How exactly this specificity is achieved is still not fully understood. Although often supposed to execute a binary decision to bind or not bind at a target sequence, in much the same way that a restriction enzyme cuts or not at its restriction site, most TFs, rather than binding to a unique sequence, in reality bind with various affinities to a range of related sequences. This molecular recognition is achieved through complementary interactions between protein and DNA surfaces and their functional groups. These interactions must provide enough information both to define the binding site sequence and to discriminate authentic binding sites from a cloud of related sites that might be made accessible by thermal fluctuations (1). To capture these interactions, genome-wide prediction of TF-binding sites and their affinities (or, ideally, binding free energies) rely chiefly on quantitative models based on experimentally/empirically determined or computationally predicted binding sites. Many of these models are mechanistically agnostic, simply exploiting the statistical enrichment of sequences recovered from in vivo or in vitro binding experiments irrespective of the detailed chemistry and physics of site recognition. These quantitative models of TF binding can also be used for predicting disease-causing mutations. The simplest model of TF binding assumes that the preference for any nucleotide within a DNA binding site is independent of the nucleotides in the remaining positions. Such independent position models are typically represented by position weight matrices (PWMs), which report, for each nucleotide at every position, this nucleotide's contribution to the total TF binding affinity score (2). Although such models have been very successful, they are known to be nonperfect. In PNAS, Zhou et al. (3) show that information about DNA shape can improve TF-binding models significantly.

High-throughput experimental techniques, such as protein-binding microarrays (PBMs) (4) or high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX) (5–8), have provided opportunities for constructing more accurate but necessarily more complex models. As natural extensions of the independent contribution PWM model (see ref. 9 for a recent review), new models often assume that the contribution of any nucleotide at a given position depends not only on the identity of this nucleotide but also on the identity of one or more preceding nucleotides (10). Consequently, the algorithms to build these models often estimate the contribution of dimers, trimers, and generally *k*-mers to the total binding affinity. Unfortunately, the number of features used in such higher-order models often increases exponentially with *k* while yielding, as shown by DERAM5 challenge (11), only modest gain. At the same time, large numbers of features become problematic, especially when the experimental data used for training are not so abundant, because it increases the risk of overtraining. Therefore, it is important to zoom in on the most informative of features.

Although the complementarity between TFs and their binding sites are obviously dependent on the local features shaping the DNA molecule in 3D (12), such as the major and minor grove surfaces, DNA bending, etc. (Fig. 1), until recently there has not been much effort to include explicitly such DNA features into TF-binding prediction models. It has been assumed that that the specification of the individual bases encapsulates and captures the interaction chemistry implicitly. However, these local shape properties are sequence dependent and thus vary and can be modeled based on the linear DNA sequence information (13). This, in turn, provides the opportunity to expand binding models to include features describing the propensity to adopt a local 3D shape (Fig. 2). Zhou et al. demonstrate that the improvement obtained by extending the independent model by including such shape-describing features is comparable to the improvement gained by including the first-order dependencies (dimers). However, using shape required a significantly smaller number of additional features. The flexibility and thermal dynamics of double-stranded DNA likely ensure that closely related binding sites populate overlapping distributions of structures; to resolve these distributions requires either a theoretically justifiable chemical–structural principle or sufficiently dense data sampling to establish their relative probabilities. DNA shape provides one such principle. The shape vector itself may be considered to constrain implicitly the vectors that partially reflect electrostatics, base stacking, hydration, etc.

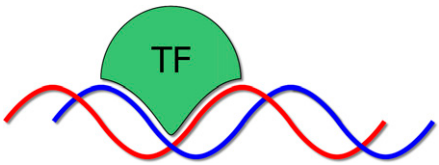


**Fig. 1.** TFs can recognize DNA shape.

An important advantage of including DNA shape among predictive features, in addition to reducing the number of features, is its biological interpretability. The propensity of a DNA fragment for a particular shape summarizes the cooperative contribution of the sequence neighborhood to that shape. A natural way for DNA shape to impact TF binding is through the free-energy differences imposed upon the double helix to conform to the shape that facilitates binding. Importantly, high-throughput in vitro measurements can also be used to model binding free energy by relating the probability of binding to interaction energies via Fermi–Dirac distribution (14, 15). Specifically, the probably of TF binding to a nucleotide fragment *S*, $p(S)$, depends on the binding free energy $E(S)$ and the chemical potential $\mu$, which is a function of the TF's concentration:

$$p(S) = \frac{1}{e^{(E(S)-\mu)/K_B T} + 1}.$$

In the independent model, the energy $E(S)$ is assumed to be the sum of the binding energies

Author contributions: T.M.P. and D.L. wrote the paper.

The authors declare no conflict of interest.

See companion article on page 4654.

[1]To whom correspondence may be addressed. Email: levens@helix.nih.gov or przytyck@ncbi.nlm.nih.gov.

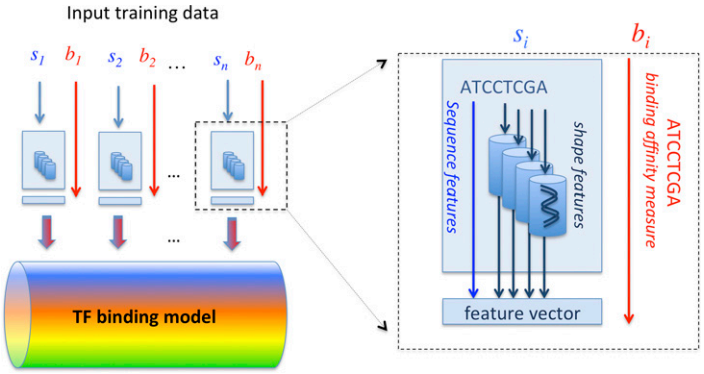


**Fig. 2.** Incorporating DNA shape features for construction of sequence based TF binding models. The model (four-color cylinder) is trained on experimentally tested sequences, $s_1, s_2 \ldots s_n$ (blue), using feature vector extracted from these sequences (blue boxes) and corresponding biding affinity measurements, $b_1, b_2 \ldots b_n$(red). The features extracted from each sequence (*Right Inset*) are of two types: sequence features and shape features. Sequence features are obtained directly from the sequence such as 1-mers, 2-mers, etc., whereas DNA shape features are obtained from each sequence by using DNAshape method (13). Each blue cylinder represents an application of a shape-type–specific (minor groove width, propeller twist, roll, or helix twist) DNAshape model to the given input sequence.

of the individual bases of *S*. A *k*-mer–based, higher-order model includes "corrections" for unspecified interactions or binding cooperativity of consecutive bases (14). Replacing such corrections with an additive contribution from DNA shape has the potential to provide a more biologically interpretable model. Such a model can potentially allow for an estimation of the contribution of shape to the energy of binding under the assumptions of this model.

Notably, Zhou et al. observe that the gain that the shape features provide over the PWM model was not uniform for all TFs. For some TF families, like bHLH, the method showed consistent improvement whereas for others, including zinc fingers (ZFs), the benefits were limited. Thus, features representing additional binding parameters might be in play. In addition, for the TFs with multiple binding domains, the sequence propensity of a given domain might be context dependent. For example, it has been found that, for ZF arrays, the binding propensities of individual fingers depend on their order (16). In vivo, additional context dependency can be imposed by changes in DNA conformation in response to transcription as exemplified by far upstream element (FUSE) binding protein (FBP) binding to the *MYC* FUSE (17), suggesting that the role of DNA topology in DNA binding goes beyond the static shape properties. In the case of FUSE, the DNA binding site comprises a single-stranded region in which DNA shape defined by sequence may become conflated with many alternative secondary structures. In instances where DNA structure is distorted by torsional or flexural stress, DNA shape as defined under relaxed conditions may become less predictive unless accounting for structural strain. Such long-range dependencies would be difficult to capture based on local sequence information only.

**ACKNOWLEDGMENTS.** D.L. is supported by Intramural Research Program of the National Institutes of Health, National Cancer Institute, Center for Cancer Research. T.M.P. is supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

**1** Schneider TD (2010) 70% efficiency of bistate molecular machines explained by information theory, high dimensional geometry and evolutionary convergence. *Nucleic Acids Res* 38(18):5995–6006.
**2** Stormo GD, Schneider TD, Gold LM (1982) Characterization of translational initiation sites in *E. coli*. *Nucleic Acids Res* 10(9): 2971–2996.
**3** Zhou T, et al. (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc Natl Acad Sci USA* 112:4654–4659.
**4** Berger MF, et al. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol* 24(11):1429–1435.
**5** Ellington AD, Szostak JW (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature* 346(6287):818–822.
**6** Tuerk C, Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249(4968):505–510.
**7** Jolma A, et al. (2013) DNA-binding specificities of human transcription factors. *Cell* 152(1-2):327–339.
**8** Liu J, Stormo GD (2005) Combining SELEX with quantitative assays to rapidly obtain accurate models of protein-DNA interactions. *Nucleic Acids Res* 33(17):e141.
**9** Stormo GD (2013) Modeling the specificity of protein-DNA interactions. *Quant Biol* 1(2):115–130.
**10** Agius P, Arvey A, Chang W, Noble WS, Leslie C (2010) High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions. *PLoS Comput Biol* 6(9): e1000916.
**11** Weirauch MT, et al.; DREAM5 Consortium (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat Biotechnol* 31(2):126–134.
**12** Rohs R, et al. (2009) The role of DNA shape in protein-DNA recognition. *Nature* 461(7268):1248–1253.
**13** Zhou T, et al. (2013) DNAshape: A method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic acids Res* 41(Web Server issue):W56–W62.
**14** Djordjevic M, Sengupta AM, Shraiman BI (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res* 13(11):2381–2390.
**15** Zhao Y, Granas D, Stormo GD (2009) Inferring binding energies from selected binding sites. *PLoS Comput Biol* 5(12):e1000590.
**16** Persikov AV, et al. (2015) A systematic survey of the Cys2His2 zinc finger DNA-binding landscape. *Nucleic Acids Res* 43(3):1965–1984.
**17** Kouzine F, Liu J, Sanford S, Chung HJ, Levens D (2004) The dynamic response of upstream DNA to transcription-generated torsional stress. *Nat Struct Mol Biol* 11(11):1092–1100.