Title: Genome analysis based on the integration of DNA sequence and shape

PI: Rohs, Remo (USC); Co-I: Noble, William Stafford (UW); Co-I: Tullius, Thomas D. (BU)

PROJECT SUMMARY

A battery of tools for genome analysis will be developed by integrating information derived from DNA sequence and its three-dimensional structure. Currently, genome analysis is mainly based on the one-dimensional DNA sequence, comprised of the letters A, C, G, and T. However, proteins recognize DNA as a three-dimensional object. Local nuances in DNA shape have been demonstrated to play a crucial role in the binding specificity of transcription factors (TFs), including those involved in embryonic development and human cancer. The basis for the novel tools developed in this proposal is a high-throughput (HT) method for the prediction of local DNA shape at genomic scale. This tool will predict multiple features of DNA shape and make the data available to the community in the data format of the UCSC Genome Browser through a user-friendly web server interface. It will be possible to easily analyze the shape of any number or length of DNA sequences, including whole genomes. The HT shape prediction will be experimentally validated based on X-ray crystallography, NMR spectroscopy, and hydroxyl radical cleavage data. The predictions will be combined with ORChID, an ENCODE project that infers DNA minor groove geometry from hydroxyl radical cleavage experiments. The HT method will be used to study the question of how paralogous TFs select different target sites in vivo while they share core-binding motifs or have similar binding properties in vitro. To answer this question, we will annotate the effect of flanking sequences on multiple structural features of TF binding sites (TFBSs). An initial focus of this study will be homeodomains and basic helix-loop-helix (bHLH) TFs before we include other protein families and construct a comprehensive TFBS database that provides shape features for binding motifs derived from JASPAR and other databases. Structural effects of single nucleotide polymorphisms (SNPs) will also be analyzed. Some SNPs are associated with deleterious function, while others have no apparent effect. The HT shape prediction method will be used to classify the impact of SNPs on DNA shape in the various genomes. Quantitative effects of SNPs on DNA structure will be correlated with expression quantitative trait loci (eQTLs) and genome-wide association study (GWAS) signals, to develop a predictive tool for the biological effect of specific SNPs. HT prediction will be used to design DNA sequences with different AT vs. GC contents but similar shape. The relative contributions of sequence vs. shape will be tested with analytic models including multivariate linear regression and support vector regression (SVR). For systems, for which the integration of sequence and shape proves advantageous, we will develop novel motif finding tools based on an extended alphabet that combines the four letters A, C, G, and T with specific structural features selected based on their significance by machine learning approaches. Sequence+shape motifs will be tested, compared to sequence-only and shape-only motifs, and integrated into the MEME Suite. The integration of sequence and shape will be achieved with the goal to increase the accuracy of finding TFBSs based on in vivo DNA binding data.

Comment [TZ1]: Relative contribution to binding affinity/specificity?

Comment [TZ2]: With specific informative structural features selected by machine learning approached?

PROJECT NARRATIVE

Protein-DNA recognition is a critical yet poorly understood component of gene regulation. This proposal will connect the fields of DNA sequence and structure analysis, which have developed along parallel lines, largely disconnected from each other. The integration of 1D sequence at genome-wide scale with 3D structure at atomic resolution will lead to the development of novel genome analysis tools and further our understanding of genome function, resulting in fundamental insights into the mechanisms of gene regulation and its impact on human disease.
We thank the reviewers for their valuable comments, which greatly helped this revision. The reviewers acknowledged that the role of DNA shape in protein binding has not been well studied to date and that the proposed studies are innovative and could have a high impact on the study of protein-DNA interactions. The reviewers assessed the application as showing much thought and attention to detail. The main critique raised by the reviewers was the need to more rigorously evaluate and validate the improvement achieved by the proposed approaches relative to existing sequence-based methods. Also, the reviewers were not convinced that the proposed approach would lead to better evolutionary insights than sequence alone.

To address these concerns, we had made the following essential changes to the revised application:

1. **We have performed additional experimental validation of the shape prediction method (Aim 1).** We found good agreement of our predictions with all available crystal and NMR structures. We performed additional hydroxyl radical cleavage experiments and found good agreement between experimental measurements and our DNA shape predictions. We also showed, using the FIS protein as a model system, that shape prediction and hydroxyl radical cleavage are predictive of its binding affinity varying over three orders of magnitude as a function of shape in a non-contacted region.

2. **We removed Aim 2 (evolutionary studies) and replaced it with an aim studying the effect of flanking sequences of transcription factor binding sites (TFBSs) on the binding specificity of paralogous TFs.** This question is both biologically highly significant and much more closely related to the core of the proposal. For preliminary data, we studied paralogous basic helix-loop-helix (bHLH) TFs that share a binding motif but bind to distinct sites in vivo. Based on our proposed approach, we described variants in DNA shape induced by different genomic flanks that are likely recognized by paralogous TF sites to select their distinct in vivo sites.

3. **We demonstrated that DNA shape explains mechanistically, to a great extent, why low-frequency SNPs in TFBSs are deleterious (Aim 3).** Using SNP frequency data from Drosophila individuals and ChIP-seq data for cis-regulatory modules, we showed that for SNPs located within TFBSs and their direct flanks DNA shape is more affected by low-frequency SNPs compared to high-frequency SNPs, which are likely tolerated. This observation finding is statistically significant and cannot be observed for SNPs located elsewhere treated as negative control for TFBSs and their direct flanks but not for a negative control group of SNPs that are not located in TFBSs. Chi-squared and permutation test indicated that a statistical analysis of the sequence composition of the mutations alone could not reveal explain this effect finding.

4. **We proposed analytical models for assessing relative contributions of DNA sequence vs. shape, which are the basis for direct tests of the hypothesis that shape leads to improved methods.** Those analytical models will be used to test the role of DNA shape on a case-by-case basis. We expect that in some cases sequence or position weight matrices (PWMs) describe binding specificity well, whereas in other cases DNA shape provides important additional information. Understanding the contributions of DNA shape to protein-DNA interactions and TF binding will provide mechanistic insights that are not apparent based on sequence.

5. **We directly tested the hypothesis that considering DNA shape leads to improved methods (Aim 4).** We used multivariate linear regression and support vector regression (SVR) to explicitly assess the hypothesis that integrating DNA sequence and shape can more accurately describe protein-DNA interactions than sequence alone. We used DNase I as a model system, and mean squared errors and R² values describing the correlation between predicted and observed DNase I cut rates as quantitative measures in the assessment of sequence-shape, sequence only, and shape only approaches. We found that the improvement of integrating shape is significant. Aim 4 now focuses on deriving analytic models for testing the contribution of shape and uses machine learning (SVR) to determine which structural features are important for a given DNA binding protein. This approach will facilitate the necessary tests preceding the integration of structural parameters into the MEME Suite.

A vertical line in the left margin highlights essential changes in the text of the application. Figures 4, 5, 7, and 11 are new and show additional data. Figures 6, 9, and 10 include additional or revised panels with new data.

Additional experimental validations are based on multiple collaborations, which will include the testing of designed promoters with identical shape but various AT/GC contents in Drosophila and yeast, X-ray crystallography, EPR, PBM, SELEX-seq, DNase I footprints (see letters of support) and other approaches. While some of those experiments are listed as alternative approaches, we will pursue all of these opportunities.

**Other progress made by the PI since the submission of the original application** includes the submission of six manuscripts for publication. Two of those manuscripts have meanwhile been published (see biosketch). A study that is very relevant to the application, as it explains the relationship between DNA shape, CpG methylation, and DNase I cleavage, was positively reviewed by PNAS (direct submission) and a revision was...
submitted. The same applies to a manuscript revised for *Cell Reports* in which we establish the role of histidines in shape readout. A manuscript on the impact of DNA shape on the binding of two paralogous bHLH TFs due to flanking sequences of TFBSs was presented at *RECOMB* and submitted to a high impact journal.
SPECIFIC AIMS

The advent of massive, genome-wide sequencing technologies in combination with computational algorithms for data analysis have provided deep insights into fundamental processes in molecular biology, genetics, human disease, and development. Currently, many current approaches to genome analysis treat the genome as a sequence of letters, representing the four nucleotides A, C, G, and T. Those letters are often used to describe transcription factor binding sites (TFBSs) in the form of motifs. Sequence-based approaches explain many characteristics of protein-DNA binding, but mechanisms leading to binding specificity are not yet well understood. As an example, it is known that some transcription factors (TFs) bind only a subset of their binding motifs and how paralogous TFs achieve distinct functions by binding to different genomic sites. We believe that the three-dimensional (3D) structure of DNA, or "DNA shape," can answer some of those questions because 3D properties are influenced by nucleotides in the flanking regions of TF binding motifs. Although nucleotide sequence is the main determinant of DNA shape, there exists a degeneracy of the sequence-structure relationship. Different sequences can lead to the same structure or, in turn, single variations in sequence can change DNA shape in an extended region of several base pairs. We have previously demonstrated that nuances in DNA shape play an important role in protein-DNA recognition, in particular in mechanisms through which TFs achieve binding specificity. With this project, we aim to develop a battery of new tools for genome analysis that connect sequence with additional information embodied in DNA shape. These approaches will reveal currently unknown mechanisms of TF binding and genome function.

Aim 1. Predict DNA shape on a genome-wide basis with a high-throughput approach. Information on DNA shape derived from experimental methods is currently limited to only a fraction of the genome. To provide such information on a genome-wide scale, we will develop a method for the high-throughput (HT) prediction of DNA shape. Our approach builds on our recently published method to predict DNA minor groove width. We will extend this approach to quantify a variety of different features of DNA shape and its conformational flexibility. We will provide HT predictions of DNA shape features in the UCSC Genome Browser and correlate them with quantitative measures of protein-DNA binding. We will develop a comprehensive database with DNA shape features for all known TFBSs derived from JASPAR. We will use data from X-ray crystallography and NMR spectroscopy to validate our HT predictions. Co-I Tullius will perform hydroxyl radical cleavage measurements for further validation and integrate the HT method with the ORChID server.

Aim 2. Annotate structural relationships between DNA targets of paralogous TFs. Closely related TFs often share a DNA binding motif and therefore appear to recognize the same motif in vitro. However, they bind to different DNA target sites in vivo, which is often the basis for executing different functions. We will use the HT method to infer the shapes of DNA binding sites of paralogous TFs in their genomic contexts. Our hypothesis is that flanking sequences that are not directly contacted by the protein induce distinct structural features that affect TF binding specificity. We will classify selected DNA shape features and use the homeodomain (HD) and basic helix-loop-helix (bHLH) protein families as pilot studies. We will compare DNA structural features within and between TF families by analyzing available high-throughput data from protein binding microarray (PBM), bacterial-one-hybrid (B1H), SELEX-seq, and DNase I footprinting experiments.

Aim 3. Characterize the effect of SNPs on DNA shape. Single nucleotide polymorphisms (SNPs) have a broad spectrum of impacts on DNA shape, ranging from no impact to dramatic alterations in shape. We will quantify the effect of SNPs on DNA shape, in order to investigate whether SNPs that cause large structural variations are more likely to have functional consequences. We will correlate the quantitative structural effects of SNPs with data on human disease and expression quantitative trait loci (eQTLs) for Drosophila and yeast. We will determine whether DNA shape can be used to classify genetic variations whose effects cannot be differentiated in sequence. If SNPs can have a broad spectrum of effects on shape, different sequences can share the same shape. This suggests a reverse approach of engineering specificity by altering base pairs not involved in direct protein contacts. We will identify such sites and test if they are selectively advantageous.

Aim 4. Derive analytic models to integrate DNA sequence and shape in motif analysis. We will derive analytic models such as multivariate linear regression to determine the relative contributions of shape and sequence and of diverse DNA shape features (e.g., minor groove width or roll angle) to binding. We will use support vector regression (SVR) models to assess the predictive power of binding affinity or DNase I cut rate when DNA sequence and shape are combined, in comparison to using sequence or shape in isolation. We will also develop new approaches using shape-augmented motif descriptions in scanning the genome for known TFBS motifs. Co-I Noble will integrate sequence-shape motifs into the MEME Suite of motif-based sequence analysis tools. We will analyze in vitro and in vivo sequence data, including ChIP-seq data for TFBSs, to benchmark the improvement of motif scanning algorithms when sequence and shape are combined.
RESEARCH STRATEGY

1. SIGNIFICANCE

Our current understanding of genome function is largely the result of high-throughput analysis of sequence-based data in the form of a one-dimensional string of letters. Nucleotide sequence describes many aspects of the genome and reveals first-order determinants of protein-DNA binding specificity (1). However, proteins recognize DNA as a three-dimensional object and the description of protein-DNA readout mechanisms can only be based on structure (2). DNA structure is a function of DNA sequence, but the sequence-structure relationship is highly complex and degenerate. The field of local DNA shape in the genomic context is understudied to date. Thus, five main points summarize the significance of this proposal:

(i) **The sequence-structure degeneracy has not been systematically studied yet for DNA.** It is well established that diverse amino acid sequences can form similar protein structures or, in contrast, that single amino acid mutations can change a protein structure dramatically. For DNA, however, the double helix was considered to be a model that sufficiently describes its structural properties. More recently, high-resolution X-ray structures and data on sequence-dependent variations of DNA shape have shown that the Watson-Crick double helix (3) is a simplified template. My own work demonstrated that nucleotides in DNA shape (e.g., minor groove width or bending) affect the binding specificity of transcription factors (TFs) involved in development (4) and cancer (5), and affect the DNA binding properties of many protein families including histones (6).

(ii) **The understanding of protein-DNA recognition at the mechanistic level requires structural information.** Proteins can form direct contacts with the functional groups of the nucleotide bases, through hydrogen bonds or hydrophobic interactions, to read the DNA sequence. The four possible base pairs (bps) can be distinguished based on their pattern of functional groups in the major groove (Figure 1A). In the minor groove, however, direct contacts cannot distinguish A/T and T/A bps, or G/C and C/G bps, because of degeneracy in the pattern of functional groups in the minor groove (7). As we showed for Hox proteins (4), many proteins instead recognize the intrinsic shape of the minor groove (narrow or wide; Figure 1B) and achieve in vivo binding specificity by recognizing shape-dependent variations in electrostatic potential (6).

(iii) **DNA shape can explain why flanking regions of TF binding sites (TFBSs) contribute to specificity.** Regions that are not contacted by the protein, can contribute to binding specificity through their impact on DNA shape in the core motif region although they are not part of the binding motif. Flanking regions include nucleotides of at least half a helical turn 5' and 3' to TFBSs, thus a total of additional ten bps (Figure 1C) (8). **Spacers between half sites** of multimeric TFs can play a similar role, as we demonstrated for the spacer-dependent binding affinity of the papillomavirus E2 protein (8).

(vi) **DNA shape takes implicitly into account any interdependencies between nucleotides within a binding motif.** Position weight matrices (PWMs) are an elegant description of first-order binding specificity. They consider the positions within a binding site contribute to binding affinity independently from other positions in the motif (11). Stormo and coworkers have recently demonstrated that considering interdependencies between adjacent positions, by using dinucleotides, increases the accuracy of a model for binding specificity (12). Our approach builds on this observation, as structure integrates complex interdependencies between all positions of at least a helical turn (10 bps). **Critically, the concept of DNA shape achieves this integration implicitly without an explicit knowledge of individual interdependencies.** In addition, DNA shape provides a compelling mechanistic explanation of why dinucleotides at certain positions increase the accuracy of motif descriptions (12).

(v) **Expanding the standard sequence-centered approach for genome analysis to 3D (or 4D including conformational flexibility) will reveal currently unavailable information that is encoded in the genome.** It will reveal information on nucleotide preferences in regions of the TFBS that are not directly contacted by the protein (13). This information can include
preferences for bending, kinking, and unwinding, all of which will contribute to a mechanistic understanding of TF binding. A genome-wide DNA shape prediction method is the basis for accomplishing this goal. We will predict multiple features of DNA shape and make the data available in the UCSC Genome Browser format (Figure 1D). High-throughput DNA shape analysis on a genome-wide basis will enhance our understanding of TF binding, and thus of gene regulation in general.

2. INNOVATION

Adding DNA shape to genome analysis will lead to a variety of future applications. Making structural aspects of DNA a component of genome analysis will be a significant advance over existing methodologies. However, local structural features of DNA cannot yet be analyzed on a high-throughput basis. High-resolution structure determination methods, such as X-ray crystallography and NMR spectroscopy are limited to short DNA segments. Molecular simulations can provide detailed structural information but are also restricted to a relatively small number of short DNA molecules (14). We propose several key innovations:

(i) The first innovation is to develop the methodology for the prediction of DNA shape on a genome-wide basis without the need for time-consuming molecular simulations. With current molecular simulation approaches, DNA shape prediction of a single short stretch of 12-18 base pairs (bps) requires a computation time of 10-14 days on a single processor (14, 15). We will predict DNA shape on a genomic scale by mining structural data that we generated via Monte Carlo (MC) predictions (4, 8, 14, 16, 17) of a representative set of DNA fragments. Our high-throughput (HT) approach will use a sliding window or k-mer approach to mine structural information obtained from those MC predictions. The resulting library of average conformations on the 512 unique pentamers that can be formed in DNA will be used to infer shape parameters at the central bp. The HT method predicts structural features for the whole yeast genome in only 20 seconds on a single processor. Through parallelization, we will further improve this benchmark, but the ability to replace days by seconds demonstrates the feasibility and innovative potential of the proposed work.

(ii) Based on whole-genome maps of DNA shape at single-nucleotide resolution we will develop a battery of innovative tools for genome analysis. As a key innovation, we will make shape information easily accessible to the research community and provide a publicly accessible prediction server, design tracks in the UCSC Genome Browser format to display DNA structural features (Figure 1D), such as minor and major groove widths, helical parameters, bending, and conformational flexibility. In addition, we will develop a comprehensive database to provide structural features of all TFBS from databases such as JASPAR (18).

(iii) The effect of SNPs in protein coding regions is fairly well understood (19, 20). However, this is not the case for non-coding regions where SNPs can either have a deleterious effect or be tolerated. Although, by definition, a SNP only changes a single nucleotide, the effect of a SNP on DNA shape can vary between a dramatic change of local DNA structure or no effect on shape. We will provide a new perspective on genetic variations, quantify the impact of SNPs on DNA shape, and associate this information with function.

(iv) High-throughput DNA shape prediction enables the integration of DNA shape with sequence-based genome analysis tools such as motif finding algorithms. We will use multivariate linear regression to implement an analytic framework for assessing the relative contributions of various DNA shape features, and sequence, to protein-DNA binding. In addition, we will use support vector regression (SVR) to combine sequence and shape parameters in predicting quantitative measures of binding. For motif finding, we will expand the four-letter alphabet of A, C, G, and T to a larger number of letters that encode the identity of the nucleotide along with structural features. The development of a motif finding algorithm based on shape-augmented motifs will provide an innovative tool for the analysis of ChIP-seq and other sequencing data.

3. APPROACH

3.1. Summary of the PI’s previous work related to the proposal. We demonstrated that DNA shape plays an important role in protein-DNA recognition (6). The first example, for which we identified nuances in minor groove shape that contribute to binding specificity, is the TF Sex combs reduced (Scr), a member of the Drosophila Hox family (4). N-terminal Scr residues recognize the intrinsic shape and electrostatic potential of the minor groove. In particular, the Scr specific site fkh250 is recognized by a set of N-terminal residues, His-12, Arg3, and Arg5 (Figure 2A) while two of those residues are disordered in the presence of a Hox consensus site (Figure 2B). In collaboration with Richard Mann (Columbia), we confirmed the functional role of Arg3 and His-12 based on in vitro and in vivo mutagenesis experiments (4). These two residues recognize a second minimum in groove width that is only present in the Scr in vivo site fkh250 (Figure 2C-D). MC and HT predictions suggest that the distinct shape of fkh250 is already present in the absence of the protein. This observation reveals intrinsic DNA shape as an origin of TF binding specificity (Figure 2E-F). We and other authors (8, 21-24), have previously made related observations, denoted as “indirect readout,” but we
discovered, for the first time, that nuances in DNA shape lead to variations in electrostatic potential with functional implications (4).

Subsequently, we generalized this finding based on the analysis of all available crystal structures of protein-DNA complexes and showed that the readout of DNA shape and electrostatic potential is key for minor groove contacts across many protein families (6). Based on this finding, we published a comprehensive review and suggested a new classification scheme of protein-DNA readout modes (2). My subsequent work has emphasized the importance of DNA shape for specific TFs (5, 25) and nucleosomes (26). However, the bottleneck for studying effects of DNA shape in protein-DNA recognition on a genome-wide basis is the incomplete coverage of experimental data and the infeasibility of generating genome-wide structural data via molecular simulations (14). Our preliminary work for this proposal overcomes this bottleneck by developing a HT analysis tool that uses a data mining approach for the genome-wide analysis of minor groove width. A first version of this method has been published in Cell and applied to the analysis of Hox binding sites derived from SELEX-seq experiments (25) (Figure 2G). Interestingly, the HT analysis reveals that anterior and posterior Hox proteins select binding sites with two minima vs. one minimum in minor groove width (25), in agreement with the only available crystal structures for Scr (4) and Ultrabithorax (Ubx) (27).

I am a New and Early Stage Investigator and have significantly contributed to the field of DNA shape analysis. Several of my papers published in Cell (4, 25) and Nature (5, 6) were highlighted in Cell (28), Nature (29), Nature Reviews Genetics (30), Nature Struct. Mol. Biology (31, 32), C&EN, HHMI, USC, and Columbia News. Our recent Cell paper (25) was selected by RECOMB and ISMB as one of the 10 most influential papers in 2011 in the fields of systems biology and regulatory genomics. A textbook (33) adopted our new classification of readout modes (2).

## 3.2. Aim 1. Predict DNA shape on a genome-wide basis with a high-throughput approach.

**Background and preliminary results.** Based on our work that shows the important biological role of the structure of unbound DNA, we developed a high-throughput (HT) analysis tool for its systematic study. To our knowledge, such a tool to predict intrinsic DNA shape on a genome-wide basis is not available yet. We will develop and validate this tool, and make it accessible to the research community. DNA shape includes multiple structural features of the double helix (e.g. minor groove width). We previously demonstrated the important role of minor groove width in protein-DNA recognition (4, 6, 25) and, therefore, focus the preliminary work on this parameter. We emphasize, though, that our HT approach can be applied to any number of DNA shape parameters.

We tested the feasibility of this approach based on a preliminary HT method for the prediction of minor groove width (25). HT predictions for Hox protein binding sites are in good agreement with computationally expensive MC predictions. The excellent agreement between MC and HT predictions is indicated by Pearson correlations of $R^2=0.89$ for fkh250 (Figure 2E) and $R^2=0.95$ for fkh250**min** (Figure 2F), which take all positions of the binding site into account. In addition, we compared minor groove widths predicted by MC and HT methods for six TFs, for which we previously established the importance of minor groove readout (6). Pearson correlations between MC and HT predictions are in the range of $R^2=0.83-0.96$ (Figure 3).

Beyond TFs, DNA shape also plays a role in regulating the strength of less specific protein-DNA interactions. Examples are the nucleosome (6, 34) and the nucleoid-associated factor of inversion stimulation (FIS) (22). We compared FIS binding sites with binding affinities differing by three orders of magnitude. In the
unbound state, the high-affinity site already assumes a minor groove geometry similar to the one observed in the complex (Figure 4A) whereas the low-affinity site needs to be deformed upon binding (Figure 4B). As a consequence, the predicted minor groove width of the five central nucleotides correlates with the logarithmic binding affinity log(Kd) for a set of seven binding sites (22) with R²=0.65. When one sequence with a central TpA “hinge” step is excluded due to its flexibility, the Pearson correlation is even stronger with R²=0.99 (Figure 4C). Since hydroxyl radical (OH) cleavage is an experimental measure of minor groove geometry in solution (35), we tested the hypothesis that, if FIS binding affinity is regulated through minor groove geometry, OH cleavage intensity should also correlate with binding affinity. Based on ORChID2 data (36), we found OH cleavage intensities to significantly correlate with log(Kd) with R²=0.83 (Figure 4D).

We previously validated the MC method that underlies our HT predictions based on X-ray, NMR, and hydroxyl radical cleavage data (4, 8, 14, 16, 25, 37). Here, we will further expand our validation approaches. To this end, we measured OH cleavage profiles for three TFBSs to demonstrate the feasibility of our plan to measure the profile of 100 additional TFBS in various sequence environments. We found for the binding sites of the phage 434 repressor (38), MogR gene regulator (39), and Msx-1 homeodomain (40) a good agreement of R²=0.57-0.85 between minor groove width and experimental OH intensities (Figure 5A-C). Furthermore, we validated the HT method for yeast in vivo nucleosome binding sites with OH cleavage data. Based on an average of 23,076 sites (41), the HT method reveals a 10-bp periodicity of minor groove width for unbound sites. The Spearman rank’s correlation with ORChID2 data (36), which also shows the 10-bp periodicity, is R²=0.80 (Figure 5D). Thus, DNA shape shows the 10-bp periodicity known from dinucleotide analysis (42, 43).

Besides these encouraging experimental agreements, rigorous validation of HT predictions with all available X-ray and NMR structures in the Protein Data Bank (PDB) provides further confidence about the selected approach. To this end, we performed a validation for the predictions of two important structural features, namely minor groove width and roll (4, 8). We used bound DNAs solved by (i) X-ray crystallography (total number of 1,640 structures) and (ii) NMR spectroscopy (101 structures). Only structures with at least one helical turn (10 bp), no chemical modifications or drastic deformations were included in this validation group. For minor groove width, we achieve Spearman’s rank correlations between prediction and experiment of 0.67 for X-ray structures and of 0.68 for NMR structures (Figure 6A). The respective numbers for Roll are 0.63 and 0.68 (Figure 6A), respectively. It should be noted that a small dataset of unbound DNAs has not been included because crystal structures of DNAs not bound to a protein are known to undergo large deformations upon crystal packing in comparison to bound DNAs. The agreement in minor groove width is nevertheless acceptable with a Spearman rank correlation of 0.57. A dataset of unbound DNAs solved by NMR spectroscopy was too small for statistical analysis but the Spearman’s rank correlation is 0.99 for both minor groove width and roll for 10 structures (14, 36, 44).
Besides those comparisons with experimental data, we also tested if a pentamer is the appropriate size of a “sliding window” (Figure 6B) for mining MC trajectories, thus replacing the MC method with a HT approach. We performed 10-fold cross-validation, by iteratively removing a randomly selected test set of 10% of the MC trajectories, and by using the remaining data for HT predictions of minor groove width for the test sequences. We calculated Pearson correlations between MC and HT predictions based on all nucleotide positions in the test sequences. The median Pearson correlation over 10 rounds of cross-validation is on average 0.90 demonstrating that a pentamer approach captures accurately DNA shape features (Figure 6C).

**Approach. Step 1. Develop methodology for calculating structural features of DNA.** We now generalize the sliding pentamere approach to predict many different structural features of DNA, which we will combine in a so-called “structural feature vector”. We will calculate those parameters based on the pentamer model using 2 million cycles of MC simulations of each DNA duplex structures that contain representations of all 512 possible pentamers. Each cycle consists of randomly varying all degrees of freedom of a finite DNA duplex as described elsewhere (8). We will record a snapshot every ten MC cycles of the remaining trajectory to generate an ensemble of 150,000 all-atom structures for each duplex. Random conformational transitions from state \( m \) to \( n \) are accepted with a probability

\[
P(m \rightarrow n) = \min \left( 1, \frac{J(n)e^{-\beta E(n)}}{J(m)e^{-\beta E(m)}} \right)
\]

based on the Metropolis-Boltzmann criterion (45) with associated Jacobians \( J \), which we previously derived (17). The calculation of total energy \( E \) is based on the AMBER force field (46) and an implicit solvent model (47).

We generated a database of MC trajectories for 2,121 sequences of 12-27 bps in length, and we will further increase this number. For each of those trajectories, we will remove the initial 500,000 MC cycles, which we consider as equilibration (8). Using those data, we will predict a set of average structural features at the central position of each pentamer. Naturally, the positions adjacent to the center will use other pentamers with their own structural features. The calculation of each feature will be carried out based on the respective definition of the parameter. Minor groove width \( w \), for example, is calculated by averaging groove width over three levels \( \{-1, 0, +1\} \) surrounding the plane of a given bp, thus including 5' and 3' inter-bp values. The value for a given pentamer \( k \) will then be determined by averaging over all occurrences \( N \) of a given pentamer in our database, thus

\[
w_i = \frac{1}{3N} \sum_{j=1}^{N} w_{i,j}
\]

Thus, each pentamer will determine the minor groove width \( w_i \) centered around bp \( i \) (colored bps in pentamers in Figure 6B).

We will use a pentamer as a sliding window to mine the MC data and determine the average conformation of each unique pentamer (Figure 6B). While this approach has been validated, the choice of the pentamer model is a
balance between accuracy and statistical coverage. In our dataset, each of the 512 unique pentamers lead to
is represented on average in \( \lambda = 44 \) different DNA fragments. The sequence environment for each pentamer
occurrence differs so that long-range effects are taken into account. The pentamer approach is based on the
approximation that nearest and next-nearest neighbor interactions are the dominating effects. In comparison, a
tetramer model does not include the effect of next-nearest neighbors, and a hexamer model would have
incomplete coverage. To increase accuracy, we will (i) perform additional MC runs to reach full hexamer
coverage, (ii) further validate the method based on experimental data, and (iii) parallelize the MC code.

**Step 2: Develop a structural feature vector as a description of DNA shape.** We will derive
numerous structural features based on the HT approach by analyzing the underlying MC predictions with
CURVES (48, 49). At the initial stage, those parameters will include (i) minor groove width, (ii) major groove
width, (iii) propeller twist, (iv) slide, (v) roll, and (vi) helix twist. We will test each of the six parameters for their
structural significance in terms of sequence dependence and use the parameters with significant information
content to generate a structural feature vector. The various shape parameters will be derived from CURVES
analyses of each of the already generated 318 million snapshots (2,121 trajectories with 150,000 snapshots
each). While groove width can be defined for each bp of a pentamer, bp step parameters define translations and
rotations between two adjacent bps. Since there are two central bp steps or dinucleotides in a pentamer,
each pentamer contributes two values to the HT prediction of any bp step parameter (e.g., slide, roll, and helix
twist). In other words, a pentamer contributes one value for bp \( i \) but two values for the dinucleotides \( (i-1, i) \) and
\( (i, i+1) \). Thus, using the two central bp steps of a pentamer to calculate bp-step parameters might reduce the
accuracy of the pentamer model. Therefore, we will add MC trajectories to transition to a hexamer model.

**Step 3: Integrate measures of conformational flexibility into HT predictions.** We hypothesize that standard
deviations of the DNA shape predictions are a reasonable measure of conformational flexibility. However,
standard deviations between average conformations might not be very informative, and they are
problematic because the coverage is unlikely to ever be identical for all pentamers. Therefore, we will identify
the conformation that is closest to the average conformation of a pentamer based on all occurrences, and use
the standard deviation of the MC sampling (statistics of 150,000 snapshots) to describe conformational
flexibility. Using standard deviations as a measure of flexibility, we will add it as elements of our feature vector.

**Step 4: Integrate HT predictions with UCSC Genome Browser.** For each of the predicted
parameters, which reveal distinct features important for protein-DNA recognition, we will generate custom
annotation tracks in the UCSC Genome Browser format (Figure 1D). The UCSC Genome Browser provides
the infrastructure to accomplish this goal. We will use the new “track hub” functionality of the browser to set up
our own local server for our custom annotation tracks. This approach will make it easy to share the data
between the Rohs, Noble, and Tullius labs, and after validation, to release the data for public usage.

**Benchmarks for success.** We will benchmark HT predictions based on available crystal and NMR
structures. For this purpose, a goal is to predict all-atom structures with 3DNA (50), in addition to the structural
feature vectors for genome analysis. Since only a few DNA sequences have been crystallized in their free form
or studied by NMR spectroscopy (14), this comparison will include protein-DNA complexes and will compare
DNA shape in its bound and unbound forms (51, 52). This approach makes the assumption that intrinsic
properties of DNA shape are conserved upon protein binding, which holds for interactions that are not
associated with major deformations. For many TF-DNA complexes, the minor groove width derived from
crystal structures shows the same general tendency as observed in their unbound binding sites. However, X-
ray data is affected by crystal packing and can result in minor groove widening beyond the range of 3-7 Å
observed in solution state DNA (2). Therefore, the minima-maxima pattern is a more important feature than the
actual values, and such essential patterns can be compared using Spearman’s rank correlation.

To achieve a large-scale experimental validation for DNA structure in solution, we will compare HT
predictions with hydroxyl radical (OH) cleavage intensities, which measures the solvent accessibility of sugar
hydrogen atoms (53, 54). Hydroxyl radical cleavage patterns thus provide information on the sequence-
dependent variations of the DNA backbone conformation in solution. We have shown in a recent publication
that OH cleavage intensity and minor groove width are highly correlated (37). Co-I Tullius has constructed a
database, ORChID (OH Radical Cleavage Intensity Database), of experimental OH cleavage intensities (35).
ORChID is the largest collection of experimental data on DNA structure in solution. I collaborated with co-I
Tullius on the development of ORChID2 (36) as an extension of ORChID that takes the double-helical topology
of DNA into account. We will, as part of this proposal, significantly increase the coverage of the ORChID
database (35). For this purpose we will perform hydroxyl radical cleavage experiments on 100 TFBSs, which
will be selected based on their coverage of unique pentamers. Each of those 100 TFBSs will be measured in
multiple sequence environments. Co-I Tullius developed a robotic platform to generate the experimental data. Once validated, the HT approach and ORChID will be integrated in a joint server. Such integration will strengthen both methods: HT shape predictions will be supported by experimental data, and ORChID (55) will be coupled with shape information beyond minor groove accessibility.

Alternative strategies. Preliminary results indicate that the pentamer model will yield satisfying results for HT prediction. However, in case the predictions are not satisfying for certain parameters, we will replace the pentamer by a hexamer model. This will require additional MC predictions because DNA can form 2,048 unique hexamers vs. 512 unique pentamers. The hexamer model will likely improve the prediction of bp-step parameters because the pentamer model fails to include next-nearest neighbor effects on dinucleotides. As ongoing efforts, we will generate MC data for sequences with low representation in the current dataset, expand the validation of the MC data based on X-ray and NMR data, and if necessary further refine the MC method.

As alternative strategies for solution state validation, we will pursue two approaches. One is the comparison with data that probes DNA structure by its chemical reaction with the uranyl ion (55). Another approach is based on EPR site-directed spin labeling (56), which will be pursued with P. Qin (USC; see letter of support). MC predictions and EPR data were validated based on a root mean square deviation (RMSD) of 0.7 Å for six double-electron-electron resonance (DEER) distance measurements per unbound TFBS.

3.3. Aim 2. Annotate structural relationships between DNA targets of paralogous TFs.

**Background and preliminary results.** Potential binding sites for every TF occur many times in the genome but not all of those sites will be occupied by the TF, or in other words, only a subset of the available target sites is functional. In addition, paralogous TFs often share a binding motif but execute diverse functions, and thus need to bind to distinct sites in the genome. A very important but yet unsolved question is, therefore, how TFs select functional binding sites, and how paralogous TFs differentiate between their in vivo target sites.

Our work on Drosophila Hox proteins, in collaboration with Richard Mann, Harmen Bussemaker, and Barry Honig (Columbia), has demonstrated that variations in the two central positions of the AYnnAY core motif are recognized by the TF although those bps are not directly in contact with the protein. We showed that variations in these nucleotide positions are recognized through their effect on minor groove geometry, which in turn causes variations in electrostatic potential and therefore affects binding affinity (4). Based on minor groove width predictions for >650,000 binding sites derived from SELEX-seq experiments, we showed that anterior and posterior Hox proteins select indeed different DNA shapes (Figure 2G) (25).

While Hox protein binding sites modulate shape through variations of non-contacted bps within the core motif, we were interested if flanking sequences can have the same impact and serve as an origin of differential binding specificity of paralogous TFs. In collaboration with Martha Bulik (Harvard) and Raluca Gordan (Duke; see letters of support), we analyzed PBM data for two basic helix-loop-helix (bHLH) TFs from yeast, Cbf1 and Tye7. Both bHLH factors share the E-box CAnnTG as core binding motif, and previous Protein Binding Microarray (PBM) data has revealed that both TFs preferentially bind to the specific E-box CACGTG (Figure 7A) (57). However, the overlap of target sites for Cbf1 and Tye7 in the genome is very limited (Figure 7B). We hypothesized that DNA shape effects induced by the flanking sequences of the binding sites play a key role in differentiating target sites that are bound by either Cbf1 or Tye7.

![Figure 7](image-url) **Figure 7:** (A) The S. cerevisiae TFs Cbf1 and Tye7 preferentially bind to the E-box CACGTG, although (B) both TFs bind to distinct genomic targets. (C) HT prediction of DNA shape of PBM probes ranked based on E-scores reveals that binding affinity of Cbf1 and Tye7 depends on shape variations (shown for propeller twist) as a function of flanking sequences, and that both TFs select distinct shapes. To further investigate the impact of genomic flanking sequences, we used probes that all share the same E-box CACGTG as core motif and used box plots to compare the selection of (D) propeller twist and (E) minor groove width by Cbf1 (light blue) vs. Tye7 (red). Statistically significant differences in shape are highlighted with asterisks (*) based on Mann Whitney U p-values of (D) 0.02, 0.01, 0.04, 0.02, and 1.1x10^{-8} at symmetric positions 9, 8, 7, 4 and 2, respectively, and (E) 0.03, 0.008, 8.7x10^{-7}, and 5.08x10^{-7} at positions 5, 4, 3, and 2, respectively.
Based on a custom PBM that tests binding in native genomic contexts, we ranked the probes for each TF based on their relative binding affinities. We predicted structural features for each probe and noticed variations as a function of binding affinity. For propeller twist, for example, structural changes are clearly visible between high and low affinity binding sites. In addition, the relationship between binding affinity and shape is different for the two paralogous TFs (Figure 7C). In order to separate the impact of the flanking sequences on TF-specific binding from variations within the E-box motif, we generated a subset of probes that all share the CACGTG core and predicted structural features to reveal the differences in the preferred DNA shapes selected by Cbf1 and Tye7. Shape parameters for sites selected by either Cbf1 or Tye7 indicate that both bHLH TFs select distinct structural features induced by regions flanking the E-box (Mann Whitney U P-values <0.05) (Figure 7D-E). Crystal structures of the yeast bHLH TF Pho4 and the human homolog USF in complex with DNA indicate that loop regions can be responsible for shape readout in flanking regions of the TFBS (58, 59).

**Approach:** **Step 1: Annotate DNA shape selection of paralogous homeodomain TFs.** We will start this aim by studying homeodomain binding sites for mouse, which expands our work on Drosophila (4, 25). We will calculate structural feature vectors based on *in vitro* data from SELEX-seq (25), PBM (60), and bacterial-one hybrid (B1H) experiments (61, 62). We will identify structural features that are under selection and generate shape-selection heat maps (Figure 7C). We will classify all homeodomain proteins based on shape characteristics for binding sites derived from various HT experiments (60-62). We will achieve this through statistical analysis as demonstrated for yeast bHLH factors (Figure 7). Richard Mann (Columbia) and Martha Bulyk (Harvard; see letters of support) will provide SELEX-seq and PBM data for mutations of N-terminal residues to test our classifications of shape features that are selected. Initial SELEX-seq data for Sct R3A/His-12 mutants have been generated and indicate that mutants do not recognize minor groove geometry as the wt.

**Step 2: Annotate DNA shape selection of paralogous basic helix-loop-helix (bHLH) TFs.** We will expand our analysis of PBM data for basic helix-loop-helix (bHLH) proteins from yeast to available data from other organisms. We expect that shape effects at the flanks of the binding sites will yield accurate descriptions of differential binding specificity. We will map similarities in structural features of TFBS within relevant TF families. To quantify effects of sequence vs. shape and of multiple shape parameters, we will use multivariate linear regression and other analytic models (see Aim 4).

**Step 3: Annotate DNA shape selection by other protein families.** We will analyze the role of DNA shape due to spacers between half sites of multimeric TFs. This analysis will include nuclear receptors (bind DNA as homodimers) (63) and the tumor suppressor p53 (binds DNA as tetramer or “dimer of dimers”) (5, 64).

**Step 4: Build comprehensive TFBS database for structural features.** We will develop a database of TFBSs that will provide a standard sequence motif and structural features (Figure 8). We will use a heat map representation to visualize structural features. Such a heat map can be expanded into individual tracks for each sequence provided by JASPAR (18), TRANSFAC (65), UniProbe (57), DNase I footprints (66), or other sequence data. These tracks will be ordered by relative binding affinity (Figure 7C) and provided for different structural features. In addition, we will provide one average profile combining all aligned probes or reads per structural feature (green/white heat maps in Figure 8).

As an example, a TFBS-shape database can further our understanding of protein-DNA recognition when a TF is able to bind to more than one distinct sequence. This is, in principle, possible because not every bp forms direct contacts with the protein (8). Multiple binding motifs for a single TF were suggested based on PBM analysis as “primary” and “secondary” motifs (10). To refine this observation, we hypothesized that different sequence motifs might, at least in some cases, share certain shape features recognized by the TF. We tested this hypothesis for four TFs with two motifs (10) and found that minor groove width profiles of primary and secondary motifs (shown as heat maps) can indeed be similar (Figure 8).

**Benchmarks for success.** Based on encouraging preliminary data, we expect that our structural feature vector will reveal structural features specific to paralogous TFs. To validate this hypothesis, we will test our predictions for homeodomain mutants in PBM and SELEX-
seq experiments (see letters of support by Richard Mann and Martha Bulyk), and if available for bHLH mutants. We expect that some structural features will not be very informative, and that the importance of other parameters will vary among TF families. We will identify essential structural features based on multivariate linear regression and statistical learning methods, including support vector regression (67) and LASSO (68), classify important structural relationships for each TF family and interpret them based on crystal structures.

Alternative Strategies: It may be the case that shape features are only relevant for some families of TFs. If we do not observe consistent correlation between shape selection and differential binding of paralogous TFs, we will investigate in more detail specific TF families that do exhibit this property.

3.4. Aim 3. Characterize the effect of SNPs on DNA shape.

Background and preliminary results. To the best of our knowledge, there is no tool available yet that systematically assesses the effect of SNPs on DNA structural features. Furthermore, structural changes due to SNPs have not been systematically related to human disease or functional data.

In preliminary work, we analyzed all possible 9-mers with a mutation at the central position. Using 9-mers with four choices at its central position, results in $4^4 \times 3^4 \times 9 = 266,144$ sequences. For each of those 9-mers, we predicted minor groove width from the -2 to +2 positions before and after one of the three possible mutations occurred surrounding the mutation at position 0 respectively, and calculated Then the Euclidean distances between 2 predicted minor groove width pattern at groove width for the unique six pairs of variations between the four possible sequences. We define this measure as as is used as a quantitative score for measuring structural changes caused by a specific SNP. For “shape-insensitive SNPs,” any of the mutations has little effect on minor groove shape. In this case, the quantitative score is around 0.3 Å (Figure 9A-B). In contrast, in other cases, mutations can have a dramatic effect on shape. For those “shape-sensitive” SNPs, the quantitative score is around 2.0 Å, and can be as high as 3.5 Å (Figure 9C-D).

To test this approach, we hypothesized that “shape-sensitive” SNPs can be functionally deleterious if they occur within or near TFBS, whereas “shape-insensitive” SNPs are more likely to be tolerated. We tested this hypothesis based on SNP frequency data from 66 fly individuals, provided by Sergey Nuzhdin (USC), and assigned TFBS in cis-regulatory modules from Drosophila, in collaboration with Angela DePace (Harvard; see letter of support) using ChIP-seq experiments. We define a SNP as frequent if it is observed more than 7 times among the 66 individuals. The total of 768 SNPs can then be separated into approximately equal groups of 385 high frequency SNPs and 383 low frequency SNPs. For both groups, we analyzed separately how strongly the SNPs alter minor groove width. We detected that low frequency SNPs change shape more dramatically than high frequency SNPs if they are located in known TFBS including 5 bp flanking regions (Figure 9E). The difference between both distributions was statistically significant (based on a Kolmogoroff-Smirnov (K-S) test, which yielded a P-value of 0.02761). In the negative control group, consisting of regions not located within TFBS or their direct flanks, however, the effect of SNPs on DNA shape is indistinguishable between the two groups of low and high frequency SNPs (K-S test P-value of 0.8728) (Figure 9F).

We further tested if this distinction could be observed based on sequence. For this purpose, we counted the frequency of the 12 possible mutations in the low and high frequency groups. We implemented a $\chi^2$ test (P-value of 0.1498), which shows that the distribution difference of DNA shape changes is not simply a result of differences in the sequence composition of mutations in the two groups. Moreover, we used a permutation test, where for each trial, for SNPs share
the same mutation type were randomly shuffled. Among 1,000 trials, only 26 attempts yielded a $P$-value of K-S test smaller than the $P$-value we obtained based on the original distributions using shape information (K-S test). Thus, the null hypothesis that the distribution difference of shape changes is a result of the difference in mutation type composition is rejected at the significance level of 0.026. This result emphasizes that DNA shape analysis provides a valid approach for the functional classification of genetic variations.

**Approach.** **Step 1: Characterize the extent to which different SNPs alter shape.** We will use our HT method to quantify differential effects of SNPs on DNA shape. Our preliminary data demonstrates that minor groove width alone can provide important information on functional roles of SNPs. We will generalize our approach and use instead a structural feature vector, which includes multiple structural parameters and conformational flexibility (Aim 1). We will study the quantitative changes represented by all elements of the structural feature vector not only at the location of the SNP but also in adjacent regions of the mutation.

**Step 2: Correlate structural effects of SNPs with disease data and integrate shape into genome-wide association studies (GWAS).** To our knowledge, DNA shape has not been used yet to explain human disease. GWAS signals in humans are usually based on multiple SNPs. Therefore, we will use Drosophila as a model system where the linkage disequilibrium extends to very small genomic regions, indicating that GWAS signals will not result in blocks of associated SNPs but should uncover individual SNPs (69). We will get those SNP information from Sergey Nuzhdin (USC; see letter of support). We will evaluate the contributions of those SNPs, and compare the results to subsets of sequences that are located in TFBSs. We will receive eQTLs data from S. cerevisiae strains (70) and HaploReg (71) database, and make the data available through a user-friendly web server.

**Step 3: Identify the role of shape in quantitative trait loci (QTLs).** We will study correlations between structural effects of SNPs and particular traits in model organisms. Our hypothesis is that DNA shape changes upon a mutation contribute to the cause for a particular trait. We will start with the analysis of recently published correlations between specific mutations in a 52-bp cis-regulatory element and expression profiles of known TFs (72). The authors of this study pointed out that OH cleavage data suggests that changes in DNA shape correlate with changes in expression for one of the most effective mutations (72). We will analyze SNPs in known S. cerevisiae strains (73) and various *Drosophila* genomes and annotate structural changes in the context of QTLs, and particularly expression QTLs (eQTLs), which can be experimentally tested. We will first identify all genes with cis-regulatory differences (74), then identify SNPs in promoter regions, predict structural effects of those SNPs, and compare the results to subsets of sequences that are located in TFBSs. We will receive eQTLs data from Sergey Nuzhdin (*Drosophila*) and Ian Ehrenreich (yeast; see letters of support).

**Benchmarks for success.** The ultimate benchmark of our HT approach is to design sequences with similar DNA shape but different sequence composition, which yield similar or “better” specificity in TF binding. To test the feasibility of this approach, we used the Scr specific site (Figure 2) and calculated how many other sequences share the same minor groove geometry while sharing its core motif AYnnAY selected through hydrogen bonds and other direct contacts. We found 56 sequences that share the same minor groove geometry with the Scr in vivo site (4) with a difference of <0.1 Å in Euclidean distance in minor groove width at 12 positions including the core motif. We evaluated whether these predictions make sense by assessing if those 56 sites in fact bind Scr. We had the required experimental data for this evaluation on hand in the form of SELEX-seq reads (25) and found that all “designed” targets bind Scr with a relative affinity >0.1 among all bound targets. Among those sites, which have experimentally been shown to bind to Scr, the AT content, interestingly, varies within the predicted 12-mers of the form nnnAYnnAYnn between 42-83%, which emphasizes the significance of studying the sequence-structure degeneracy of DNA.

Using this approach of designing sequences with “better” or similar shapes, we will test some of those predictions experimentally in yeast and *Drosophila* (see letters of support from Ian Ehrenreich (USC) and Angela DePace (Harvard)). In addition, we will develop analytic models (see Aim 4 for details) that quantify the individual contributions of different nucleotide positions and various structural features to the overall shape.

**Alternative strategies.** An alternative approach is that we correlate structural feature vectors with DNase I sensitivity QTLs (75), which have recently reached a high enough resolution to identify TFBSs (66, 76).

### 3.5. Aim 4. Derive analytic models to integrate DNA sequence and shape in motif analysis.

**Background and preliminary results.** To our knowledge, no existing sequence-based analysis tool makes use of DNA shape data. Based on our previous work (6), we hypothesize that taking into account both sequence and shape information will improve our ability to accurately characterize in vivo TFBS, relative to a
sequence-only approach. Specifically, we will investigate whether the integration of sequence and shape improves the description of protein binding sites. As proof-of-principle study, we investigated the DNA binding of the endonuclease DNase I (77, 78) because its relatively weak binding selectivity cannot easily be described by a sequence motif. My group, in collaboration with John Stamatoyannopoulos (UWash) and Harmen Bussemaker (Columbia), compared the minor groove widths of all 4,096 possible hexamers with their DNase I cut rates. In a region where the DNA is contacted by the protein (Figure 10A-B), we observed that DNase I cut rate correlates with minor groove width (Figure 10C). The cut rate is high when the groove is narrow at the -3 and -2 positions, where arginine side chains are bound, and wide at the -1 and +1 positions (Figure 10C), where an opening of the groove provides better access and thus enhances the cleavage process. This observation indicates the importance of DNA shape information in explaining DNase I cleavage propensity.

**Approach.** Motivated by promising preliminary results, we will develop a rigorous analytical framework to assess the relative contribution of shape compared to sequence, and of various structural features relative to each other. We will apply this analytical framework to predict response variables, such as binding affinity or, as in our model system, DNase I cleavage rate.

**Step 1: Develop analytic multivariate linear regression models for assessing sequence and shape contributions.** Using multivariate linear regression, we showed for the DNase I system that a combination of discrete nucleotide identity with continuous values for structural parameters. Here we use the hexamer model for the DNase I binding site. The response variable $R$, in this case $-\log$(cut rate), depends on (i) the sequence A, C, G, or T at N positions (described by an indicator function $i(N)$, which is 1 for a given nucleotide and 0 otherwise), (ii) the minor groove width $w$ at each bp (MGW), and (iii) the Roll angle $r$ between adjacent bps. We applied this multivariate model to the hexameric DNase I binding site and found that the mean squared error decreases by 24% when sequence and shape are combined in comparison to sequence alone (Figure 11A). In addition, the correlation weights $\alpha$ provide information about the relative contribution of sequence and shape features at each position of the binding site. We will further develop this framework and provide this approach as an analysis tool of TFBSs.
Step 2: Develop analytic support vector regression (SVR) model for integrating sequence and shape. To achieve a higher accuracy in predicting response variables that characterize binding, we will apply SVR models (79). To assess the feasibility of this approach, we trained an SVR model using a Gaussian kernel for predicting DNase I cut rate based on the integration of sequence, minor groove width, and roll in comparison to either only sequence (Figure 11B) or only shape (Figure 11C). Using 10-fold cross validation, this SVR model predicts DNase I cut rate with a higher accuracy compared to sequence alone ($R^2$ increases by 16% and mean squared error decreases by 76%). We will use the SVR model to derive relative contributions of sequence and shape by training different models using various combinations of sequence and structural features. We will also test the contribution of each structural feature at any position of a binding site by removing individual parameters or positions at a time. For this purpose, we will re-train the SVR model, and assess contributions based on changes in $R^2$ and mean squared error. We will integrate the results in our TFBS database and expect that the data will provide mechanistic insights into binding.

Step 3: Create a motif-scanning tool for sequence+shape motifs. As an initial test of our idea to combine sequence and shape in a single motif description, we modified an existing motif discovery algorithm (80), to use an 8-letter alphabet, in which we will integrate minor groove width with sequence by using lower case letters if the minor groove is narrow at a position or upper case letters if the minor groove is wide while the letters represent nucleotides. Using such a letter-case representation provides sequence and shape information while it is still readable (Figure 12). We used the original and the modified motif discovery algorithm to analyze the ChIP-seq data for the mouse Hoxa2 TF (81). We successfully identified the known Hoxa2 in vivo binding site using either alphabet. We also found that the sequence+shape motif is consistently more enriched within ChIP-seq peaks than the sequence-only motif and that the sequence+shape motif better distinguishes between candidate motifs sites within a ChIP-seq peak and similar sites outside of a peak. In addition, we analyzed the Harbison et al. ChIP-chip dataset for yeast TFs (82) and found 6 known TFBS that cannot be found based on sequence alone. These are just initial tests for the feasibility of combining sequence and shape in genome analysis tools. To this end, we used minor groove width as the only structural features, and we chose a more or less arbitrary cutoff between “wide” and “narrow” groove. We will, as a step that precedes this work, apply machine learning techniques to determine which structural features are important for a given protein family and which cutoff in a specific structural parameter best describes distinct shapes.

Step 4: Integration into the MEME Suite. If the validations of the motif scanning using sequence+shape motifs proves successful, co-I Noble will add the search for shape-augmented motifs to the MEME Suite (83). The FIMO software (84) is a core component of the MEME Suite of motif discovery tools (83). FIMO uses a standard log-likelihood ratio scoring procedure, coupled with a dynamic programming algorithm and false discovery rate multiple testing correction, to identify and assign statistical confidence estimates to occurrences of a given motif within a given set of sequences. We will modify the FIMO software to allow the user to specify arbitrary alphabet parameters (number of characters, ambiguity codes). For this purpose, we will develop TF binding
motifs that integrate sequence information with shape in a common logo representation (85).

**Benchmarks for success.** As discussed above, we will measure the relative utility of sequence-only versus sequence+shape motifs by measuring their ability to identify motifs that occur within ChIP-seq peaks. We will use a "peak-centric" gold standard (86), which defines a single position in each ChIP-seq peak as a binding site for the TF (a positive), and treats all other genomic positions as negatives. The single positive position assigned to each ChIP-seq peak is the position with the highest PWM score. We will measure discrimination quality using a precision-recall curve, which appropriately accounts for the large disparity between the number of "positive" and "negative" sites in a given genome. We will test for statistical significance using an appropriate non-parametric statistical test, aggregating over many TFs. We will use ChIP-seq and ChIP-chip datasets (82, 87), coupled with motif databases (18, 65), to test our ability to accurately characterize TF binding motifs. Specifically, we will follow our protocol (86) for evaluating a motif scanning algorithm, add shape information to known motifs and measure the ability of the motif to distinguish between candidate sites.

**Alternative strategies.** Alternatively, we will perform a clustering of the six-dimensional structural feature vectors [Aim 1], with the aim of reducing these high-dimensional shapes to a low-dimensional discrete alphabet. We will then merge this shape alphabet with the DNA alphabet, and repeat our validation experiments. Co-I Noble developed the representation of TF binding motifs, based on sequence kernels and SVR (79). As we have previously demonstrated (88), kernel methods provide a formal framework for drawing inferences from heterogeneous data types. We will define a simple vector kernel on DNA shape (e.g., a scalar product on the structural feature vectors) and combine the sequence and shape kernels via summation (89).

**3.6. Timeline of proposed research.** While Aim 1 is the basis for the other aims, Aims 2-4 do not depend on Aim 1. The necessary initial tool for predicting structural features of DNA is functional and, thus, work on all four aims can begin simultaneously in year 1 without any dependencies. For the progression of the proposed work in each aim, we will focus on completing approximately one of the outlined steps in each aim per year.

**BIBLIOGRAPHY & REFERENCES CITED**


