1. Introduction to Application

Based on the thoughtful comments and helpful suggestions provided by the reviewers of the original proposal, the proposal has been thoroughly revised and its scope has been expanded. In particular, we added a feedback loop for the integration of computational and experimental data (Fig. 2), and added a second experimental method, EPR spin labeling, for mapping major groove geometry. Due to the substantial revision of the proposal, changes are not highlighted. Figures 1, 2, 5, and 8-13 are new. The revised proposal includes new preliminary data, which include (i) the first use of the high-throughput MC prediction tool for minor groove shape analysis of 650,000 Hox TF binding sites (published in Cell, Dec 8, 2011), (ii) a metric for mapping minor groove shape based on hydroxyl radical cleavage (published in ACS Chem. Biol., Dec 16, 2011), (iii) proof that naked DNA shape can be a predictor of protein-DNA binding affinity and specificity (Fig. 8), (iv) proof that DNA shape can help in finding TF-DNA binding motifs in ChIP-seq data (Fig. 12), and (v) a direct comparison of DNA shape prediction and hydroxyl radical cleavage data for 50,000 nucleosome binding sites (Fig. 13). The new preliminary data demonstrate the feasibility and impact of the proposed work. Several new support letters confirm that the proposed development of tools for genome-wide estimation of DNA structure is timely and will likely transform the way DNA sequence information is analyzed. The most compelling justifications for whole-genome analysis are (i) the demonstrated identification of TF binding motifs that could not be identified based on sequence, and (ii) the correlation between DNA shape, Fis-DNA binding affinity, and DNase I cleavage rate.

The new preliminary data in this proposal, PI Rohs’ recent Cell paper, and PI Rohs’ and co-I Tullius’ first joint publication provide confidence that the proposed research is likely to be successful. PI Rohs showed in his Cell paper that a high-throughput (HT) DNA structure prediction method is not merely incremental to his original Monte Carlo (MC) method, but in fact permits applications that could not be attempted with the original approach. The original MC method was able to predict the shapes of tens of sequences in days, while the HT method can predict the shapes of hundreds of thousands of sequences, or whole genomes, in minutes. The Cell paper shows that both methods predict meaningful ensembles of structures that correlate well with crystal structures and in vivo and in vitro binding data. This publication also shows that Aim 2 is not at risk of Aim 1 not succeeding, since biologists have already started to use preliminary versions of the shape analysis tools.

We agree with the reviewers that DNA structure is more than minor groove width. While the importance of minor groove shape in protein-DNA recognition has been well documented, the goal of this project has always been to predict all-atom structures and flexibility, key for the binding of many TFs. To address this comment we have added an analysis of major groove geometry, both for the preliminary version of the HT server (Fig. 11), and by obtaining experimental data on major groove shape. The all-atom structural ensembles and associated standard deviations that our approach will produce will also provide information on bending and torsional flexibility, features of DNA that the PI/Co-Is have discussed in previous publications. This includes the role of proteins in inducing DNA shape changes upon binding. The proposers are well aware that protein binding can have a major effect on DNA shape. In fact, the effect of protein cofactors on protein-DNA specificity was a central feature of PI Rohs’ recent paper in Cell. However, the current proposal focuses on the role of sequence-dependent DNA shape in protein-DNA recognition, and the goal is to map DNA shape throughout entire genomes. Intrinsic shape and induced shape will, in this proposal, be distinguished by comparing available crystal and NMR structures of protein-DNA complexes with predicted shapes of unbound DNA. There are thousands of protein-DNA complexes in the PDB but fewer than 100 free DNA structures with at least one helical turn. As PI Rohs’ previous work on the Hox and p53 TFs shows, much can be learned about protein-DNA binding specificity based on the shapes of unbound TF binding sites.

With respect to reviewers' comments on computational methodology, the MC approach is published and several validation cases are reported in the literature. MC predictions provide ensembles of structures that will be individually assessed, while standard deviations will provide information on flexibility. PI Rohs has initiated a collaboration with D. Beveridge and others to compare and integrate Molecular Dynamics data (see support letters from R. Lavery and B. Jayaram). The Dixit/ Beveridge webserver was designed to predict structures of DNA binding specificity based on the shapes of unbound TF binding sites.

The budget has not been reduced but now includes one additional co-I. The USC cluster is fully available.
2. Specific Aims

This proposal is a resubmission in response to PAR-09-218 ‘Innovations in Biomedical Computational Science and Technology (R01)’. The proposed research is aimed at developing an automatic compute server for the ad-hoc high-throughput prediction of sequence-dependent DNA shape on a genomic scale, which will be freely accessible to biomedical researchers. The included 12 support letters demonstrate the strong interest of the field in having such a server. The proposal includes first applications in analyzing transcription factor (TF)-DNA binding and nucleosome formation. As a result of this analysis, we will develop a novel way of describing TF binding sites through a combination of DNA sequence and shape in sequence-shape logos, which we will derive and curate for all known TF binding sites in a publicly accessible database.

DNA sequence analysis views DNA as a linear string of letters. Yet, the nucleotides A, C, G, and T are chemical entities that give rise to the three-dimensional structure of DNA. The resulting shape of the double helix (e.g., major/minor groove shape) is sequence-dependent and plays a role in gene regulation because the geometry of both grooves affects the accessibility of base pairs for interactions with proteins. PI Rohs has shown in recent work that DNA shape can be the source of transcription factor specificity and that the strength of non-specific interactions in the nucleosome can also be regulated by DNA shape. However, structural knowledge of DNA at atomic resolution derived from experimental studies is limited. This is especially true for DNA binding sites without the protein present. In addition, high-throughput sequencing technologies continue to provide a wealth of functional genomic sequences for which structural information is unavailable.

DNA structure and sequence studies are two areas of research that have developed along parallel lines, largely disconnected from each other. While structure research provides three-dimensional information on DNA and proteins at atomic resolution, sequence research analyzes DNA in a high-throughput manner and on a genome-wide basis. Therefore, a server for computational high-throughput prediction of the sequence-dependence of three-dimensional DNA structure will be a valuable tool to understand how functional information is encoded in the genome. The goal of this proposal, ‘High-throughput prediction of DNA shape on a genomic scale’, is to reveal far-reaching biological implications of DNA shape by connecting DNA sequence and structure analysis on a genome-wide scale based on the following Specific Aims:

Aim 1. Integrate a Monte Carlo simulation algorithm for DNA shape prediction with experimental data: hydroxyl radical cleavage (minor groove maps) and EPR spin labeling (major groove maps). PI Rohs developed a Monte Carlo (MC) algorithm for all-atom DNA structure predictions, and successfully applied this method to the study of several TF binding sites. The MC method will now be integrated with experimental measurements of DNA shape. Co-I Tullius introduced hydroxyl radical (•OH) cleavage intensity measurements as an experimental tool to map minor groove geometry. Co-I Qin developed an EPR spin labeling technology for measuring distances across the major groove. MC sampling, •OH cleavage, and EPR spin labeling all provide information on three-dimensional DNA structure in solution and can, therefore, be directly compared and integrated. We will use hydroxyl cleavage and EPR spin labeling data to validate MC predictions, and we will construct a feedback loop whereby experimental measures are used to improve the MC algorithm.

Aim 2. Develop a compute server for high-throughput DNA shape prediction on genome-wide basis. The ever-growing number of whole-genome sequences makes it imperative to develop a high-throughput server for ad-hoc prediction of three-dimensional DNA shape on a genomic scale. The resulting all-atom structural information will be combined with the output of the ORChID server for genome-wide prediction of hydroxyl cleavage patterns, developed by Co-I Tullius in collaboration with PI Rohs. Whole-genome analysis of DNA shape will be important in understanding how structural properties of DNA regulate genome functions.

Aim 3. Apply high-throughput DNA shape prediction to genome-wide studies of transcription factor-DNA binding, and derive sequence-shape logos and shape heat maps. The DNA shape prediction server will be used to analyze sequencing data, with the goal of identifying structure-based determinants of TF binding specificity. To compare DNA shape with nucleotide sequence, we will combine sequence and shape information by computing sequence-shape logos for TF binding sites. To further investigate the relationship of DNA structure to protein recognition, we will develop heat maps that are based solely on DNA shape, and not on sequence. The shape heat maps translate minor and major groove width into a one-dimensional color code.

Aim 4. Study the effect of DNA shape on nucleosome formation and chromatin organization. We will use the DNA shape prediction server for the study of non-sequence-specific interactions. We will focus on the question of whether the structure of a nucleosome-occupied region in its unbound state assists in wrapping the DNA around the histone octamer. Since the genomes of various organisms have different A/T vs. G/C content, we will compare the structure of genomic DNA across species. Comparisons of the human malaria parasite’s genome (about 80% A/T) with the yeast and fly genomes will reveal functional roles of various A/T contents.
3. Research Strategy
3.a. Significance

Gene regulatory processes require highly specific interactions between transcription factors (TF) and DNA. The molecular basis of these processes can only be partially understood based on the nucleotide sequence of TF binding sites alone. The relatively short binding motifs for a given TF occur many times in the genome. It has been known for some time that taking into account the three-dimensional structure or shape of a TF binding site increases our understanding of protein-DNA recognition (1). Additional nucleotides that are not part of the TF binding motif due to the lack of direct protein contacts, affect DNA shape, and thus the accessibility of the major and minor groove and their electrostatic potential (shape or indirect readout) (2). While TFs recognize only a small number of base pairs through the formation of hydrogen bonds or hydrophobic contacts between functional groups of the bases and protein side chains (base or direct readout), other base pairs that are not contacted by the protein also contribute to protein-DNA binding specificity (3).

PI Rohs and collaborators have shown that positively charged arginine residues of DNA binding proteins recognize DNA because they are attracted by the enhanced negative electrostatic potentials in a narrow minor groove. Such contacts were observed for a number of TF families (1). For the Hox protein Sex combs reduced (Scr), the residue Arg3 contacts a narrow minor groove region and was shown by in vivo experiments to confer binding specificity (4). While crystal structures of Scr and its cofactor Extradenticle (Exd) bound either to the Scr-specific site fkh250 or to a Hox consensus site showed that the minor groove geometry is distinct for the two DNA sites, this observation does not prove that the distinct groove geometry is intrinsic to the DNA sequence. The distinct shape of the minor groove that confers specificity might actually be induced upon protein binding. Only a comparison with the naked DNA structure would permit the conclusion that Scr recognizes an inherent sequence-dependent feature of the fkh250 site. Since experimental structural data for the fkh250 site was unavailable, PI Rohs used his Monte Carlo (MC) approach to show that the narrow minor groove is indeed an intrinsic property of the Scr-specific sequence (4).

Another example is the binding of the tumor suppressor p53 to its response element. PI Rohs and colleagues showed that Arg248 of p53, the residue most frequently found to be mutated in human tumors, binds to a narrow minor groove region with enhanced negative electrostatic potential (5). Similar questions arise for many DNA-binding proteins (1). In many cases the structure of the protein-DNA complex is known, but the structure of the unbound DNA target is unavailable (6). In such cases, a computational prediction can fill this gap by identifying DNA-shape-based specificity determinants. This illustrates one important aspect of the significance of the proposed method development.

As a second and major aspect of the significance of this proposal, the wealth of whole-genome sequence information that is now available underlines the significance of developing a server that instantly computes DNA shape on a genomic scale. Such a tool will be a crucial step toward a novel approach to analyzing genomic information. The included 12 support letters from biomedical researchers stress the importance of having such a server. Nucleotide sequence is a one-dimensional view of the information content of a genome; three-dimensional structure increases the information content that is inherent in a genome. The proposed combination of shape with sequence information in sequence-shape logos and the development of shape heat maps will make DNA shape widely accessible to biomedical researchers. This approach of translating structural information back into a one-dimensional letter code will influence current sequence-analysis and motif-search algorithms (7, 8). Therefore, the main significance of this proposal is that DNA shape, which is biologically important, will become accessible to the interpretation of genomic data.

Gene regulatory processes are also affected by the formation of nucleosomes and the organization of the genome in chromatin. While the interaction of DNA with histones is mainly non-specific, the periodic distribution of DNA sequence motifs, such as dinucleotides, has been found to affect nucleosome positioning (9, 10). The same periodicity of ten base pairs has been observed for DNA structure when bound to histones (11, 12). The ability to predict DNA shape on a genomic basis is of high significance because it will enable us to test whether the structure of naked DNA that forms nucleosome binding sites reflects the periodicity that is present in the nucleotide sequence and in the structure of the DNA when it is bound to the histone octamer in a nucleosome. The availability of a high-throughput shape prediction server is of significance because it allows the comparison of the shapes of naked nucleosome binding sites among different species with various A/T contents (e.g., yeast (13, 14), Drosophila (15), and the human malaria parasite (16)).

DNA shape is of course just one factor in determining protein-DNA binding specificity, and proteins play an important role as well (17, 18). This proposal focuses on the role of DNA (1, 2, 4), because this approach will allow for genome-wide analysis of DNA shape (19), which is timely given advances in sequencing technology.
3.b. Innovation

High-throughput sequencing technologies generate a vast amount of genome-wide sequence information. The main innovation of this proposal is that we will make it possible to derive DNA shape information from existing genome sequences. The work described in this proposal will make DNA shape as accessible to biomedical researchers as nucleotide sequence is now, which will catalyze the integration of DNA shape and sequence information. We will predict DNA shape properties on a genomic scale and add this information to the UC Santa Cruz Genome Browser, as illustrated in Figure 1. The new and innovative description of TF binding sites through sequence-shape logos will change our current interpretation of genomic data and our understanding of TF-DNA binding specificity. The ability to predict DNA shape on a genomic scale will enhance our ability to understand the driving forces for nucleosome formation and chromatin organization.

PI Rohs just published a paper in Cell in which he makes the first step towards combining DNA sequence and shape analysis of thousands of Hox binding sites (19). By using a preliminary approach to infer minor groove shape, Rohs shows how DNA shape distinguishes the binding sites of the 8 Drosophila Hox proteins and reveals evolutionary relationships between Hox genes (19). This first application of a high-throughput DNA shape prediction method underlines the significance of being able to analyze the shapes of many sequences. This work for the first time predicts DNA shape for high-throughput sequencing data (19).

3.b.1. Integration of DNA structure prediction with large-scale experimental data on DNA shape in solution. Structural data for naked DNA is particularly sparse because the application of X-ray crystallography and NMR spectroscopy to free DNA has been limited, in comparison to protein-DNA complexes (6). Difficulties in the crystallographic study of naked DNA include the flexibility of DNA, which makes crystallization of DNA a challenging task (6), and the common observation that the crystallized form of DNA is deformed due to crystal packing. NMR spectroscopy requires a multitude of NOE constraints to solve the structure, which DNA lacks as it is not as condensed as globular proteins. Computational structure prediction is the innovative method of choice to fill this gap. PI Rohs developed a Monte Carlo (MC) algorithm for the efficient conformational sampling of DNA structure (3). This approach will now be integrated with large-scale experimental measurements of DNA shape features. Co-I Tullius will obtain hydroxyl radical cleavage data for free DNA in solution (20), which allows one to infer minor groove geometry. In a complementary approach, co-I Qin will measure distances across the major groove based on an EPR technique that uses nitroxide spin labels (21, 22). Both experimental measurements yield distance information, one across the minor and the other across the major groove. The MC prediction algorithm will be continuously refined in a feedback loop based on new experimental data, and the choice of experimental studies will be informed by MC data by selecting sequences for both experimental techniques.

In addition, we will compare the MC method with all available crystal and NMR structures of free and bound DNA in the Protein Data Bank (PDB) and with Molecular Dynamics (MD)-generated structural data (23) for all 136 tetranucleotides that have been produced by the Ascona B-DNA Consortium (ABC). This analysis will also be part of the feedback loop that continuously refines the MC algorithm for shape prediction (see support letters from Dr. Richard Lavery and Dr. B. Jayaram).

3.b.2. Genome-wide high-throughput DNA shape prediction. The current MC sampling technique developed by PI Rohs has led to the use of MC simulation for shape prediction of approximately 2,000 different DNA sequences. These data were used to validate the method for a number of TFs. The method has been most extensively tested for Hox binding sites, for which predictions confirm crystal structures and add new information. This work led to two Cell publications (4, 19). This extensive dataset provides coverage for all 512 unique pentanucleotides, which in turn can now be used for the development of a compute server for ad-hoc high-throughput structure prediction for any DNA sequence up to the length of entire genome. Only such an instant and easily accessible tool will make DNA shape information accessible to biomedical research. Due to
the vast amount of sequence information that is now available, this is an overdue innovation, given the recognized biological role of DNA shape (1, 4, 19, 20, 24). The shape prediction server will be the first tool to assess all-atom DNA structure on a genomic scale. The importance of having such a tool is emphasized in a number of support letters that accompany this proposal. The shape prediction server will be combined with co-I Tullius’ ORChID server (20), which provides complementary structural data related to the solvent accessibility of the DNA minor groove. The new server proposed here will produce all-atom models, allowing for detailed analysis of DNA structure (e.g., groove geometry, DNA bending, flexibility, and electrostatic potential).

3.b.3. Sequence-shape logos and shape heat maps for TF binding sites. TF binding sites are commonly described as motifs, a few base pairs long, for which the probability of occurrence of A, C, G, or T is given in a Position Weight Matrix (PWM) or sequence logo. The individual positions in a PWM are independent from each other. This model can describe the selection of nucleotides due to hydrogen bonds or other direct contacts with the protein. In terms of shape, however, individual base pairs are correlated and no longer independent from each other since structural variation is based on intra-molecular interactions between adjacent base pairs. In addition, base pairs that are not contacted by proteins can still contribute to shape. Therefore, we will introduce a novel approach for combining sequence and shape information in sequence-shape logos in a format analogous to sequence weblogs (7). We expect that these innovative logos will describe protein-DNA binding specificity more accurately than sequence information alone. In another representation, we will remove sequence information and translate minor and major groove profiles into shape heat maps to emphasize DNA structure. This information can then be used to easily detect shape variation and conservation patterns, as was done in PI Rohs’ recent Cell paper (19). Shape information will be derived from our prediction server. We will introduce this key innovation, as stressed by several of the included support letters, by establishing a publicly accessible library of sequence and sequence-shape logos and shape heat maps for all known TF binding motifs. This innovation will transform the current methodology for motif finding algorithms (8) by incorporating structural information.

3.c. Approach

High-throughput sequencing technologies require computational tools capable of analyzing vast amounts of data. This is particularly challenging if the goal is to analyze DNA shape. The data format of three-dimensional DNA structure is more complex than the letter-based sequence format. As part of this proposal we will develop computational tools for DNA shape prediction and analysis. These tools include the further improvement of PI Rohs’ Monte Carlo (MC) technique (3, 4, 19, 25, 26), the development of a compute server for high-throughput DNA shape prediction, and the combination with co-I Tullius’ ORChID server (20, 27). Our efforts will include the development of a sequence-shape logo and shape heat map database for TF binding sites and of an interface for DNA structure analysis (28) and the calculation of electrostatic potential (29, 30).

The core of the approach is a feedback loop between computational predictions and experimental data as illustrated in Figure 2. PI Rohs will use his MC method to predict all-atom structures (3, 4, 19). Co-I Tullius will measure hydroxyl cleavage intensities to infer maps of minor groove geometry (20). Co-I Qin will use the EPR spin labeling technique (21) to map major groove geometry. PI Rohs will compare his predictions with those measurements, and integrate the experimental data with the MC method. In addition, the structural results will be compared to crystal and NMR structures from the PDB and MD trajectories from the ABC consortium. This feedback loop for data integration will be used to generate a library of all 2,080 unique hexamers, forming the basis for genome-wide shape analysis.

Figure 2: Approach - Integration of computational and experimental data. For a sequence of unknown structure, the MC method will predict a three-dimensional all-atom structure. This structure will be compared with EPR spin labeling and •OH cleavage data, and in turn, the experimental data on major and minor groove shape will be used to derive additional terms for the force field, which will improve the MC predictions. MC data will be, at all stages, compared to crystal and NMR structures from the PDB and MD trajectories from the ABC consortium. This feedback loop for data integration will be used to generate a library of all 2,080 unique hexamers, forming the basis for genome-wide shape analysis.
biology. Rohs just published in Cell his first paper as co-corresponding author (19), in which he used the concept for high-throughput DNA shape prediction that is the basis for this proposal. He will be assisted by co-I Tullius (Boston University) who established in a recent paper in Science that DNA shape is under evolutionary selection (20) and co-I Qin (USC) who has developed an EPR spin labeling approach for measuring distances in DNA in solution (21, 22). Rohs and Qin are at the same campus, and Rohs and Tullius have just published their first joint paper highlighted on the cover of ACS Chem. Biol. (27), demonstrating the close collaboration that will be the hallmark of this project.

We first introduce the key established methodologies or novel components that will be used in this proposal (3.c.1 to 3.c.5), and then describe the proposed work (3.c.6 to 3.c.8) including preliminary data pertinent to each specific aim.

3.c.1. Computational approach for Monte Carlo (MC) simulations. MC simulations are the source of subsequent compute server predictions as the structural data for selected individual sequences are generated in MC simulations. The molecular model for the MC simulation uses random sampling of DNA conformation based on collective and internal variables (3, 26). A basic assumption of the model is that the bond lengths and the aromatic ring systems of the bases are rigid. This approximation allows for a significant reduction in the degrees of freedom. Figure 3a illustrates the molecular model that is used for the MC sampling (26). The model is based on twelve MC variables per nucleotide. In addition to varying the positions and orientations of nucleotides as rigid bodies and the sampling of the phosphodiester backbone, thymine methyl groups are rotated, which increases the number of MC variables by one for thymine bases. Some moves lead to strand breaks 5’ and 3’ to a moved nucleotide. A novel chain closure algorithm is used to re-establish bonds following an MC move that breaks bonds (26). Chain closure incorporates sampling of endocyclic bond and torsion angles of the phosphodiester backbone and includes associated Jacobians (26).

Applying chain closure in bond angle/torsion angle space (Figure 3b) is computationally less costly compared to other approaches in which chain closure is performed only in torsion angle space (Figure 3c) while bond angles are constrained (31-34). While numeric chain closure leads to as many as sixteen solutions, and only the nearest solution is used in most methods (33), the approach used in Rohs’ model reduces chain closure to the solution of a simple quadratic equation (26). A quadratic equation has either no solution or two solutions, which can represent two distant conformations that are energetically evaluated as part of the MC sampling. These two distant conformations can belong to energy minima separated by large energy barriers, demonstrating that MC sampling in appropriate coordinates enables the global search for energy minima regardless of barrier heights.

The combination of the molecular model and analytic chain closure leads to efficient sampling as it allows for local moves based on rigid-body translations and rotations of individual nucleotides (26). In addition, the analytic chain closure is fast and enables the sampling of non-canonical backbone conformations such as α/γ-flips and BI/BII transitions (3), and transitions from canonical to non-canonical backbone conformations are reversible. As for the MC simulation protocol, we use ideal B-DNA without any sequence-dependent structure as the starting configuration for MC simulations. Thus, nucleotide sequence is the only input for DNA shape analysis. Idealized B-DNA is built with the JUMNA algorithm (35). MC simulations with two million MC cycles are performed on the USC high-performance computing cluster. The initial 500,000 MC cycles are considered as an equilibration period (3). Analysis of MC trajectories involves calculation of structural parameters based on CURVES (28) for snapshots every 10th MC cycle. The MC ensemble is analyzed in terms of average structural parameters (e.g., helical parameters, sugar puckers, dihedral angles, groove geometry, and bend angles). Standard deviations in these parameters represent flexibility and will be described along with shape.

3.c.2. Experimental approach for mapping minor groove shape: hydroxyl radical cleavage. Co-I Tullius’ lab has developed a method that is well suited for inferring structural information of DNA in solution (36) on a high-throughput basis. This method takes advantage of the sensitivity of the hydroxyl radical to local DNA structure. The hydroxyl radical, an extremely reactive free radical, attacks deoxyribose groups and cleaves...
the DNA backbone at the site of reaction (37, 38). Co-I Tullius has shown that the extent of cleavage of a given nucleotide in duplex DNA depends on the solvent-accessible surface area (SASA) of sugar hydrogen atoms (39), in particular the ones bound to the C5’ and C4’ atoms (red and blue surfaces in Figure 4a). Since the SASA of these sugar hydrogen atoms is increased for regions with wide minor groove and reduced for regions with narrow minor groove, as illustrated in Figure 4c-d, the cleavage intensity pattern represents a topographical map of the shape of the surface of the DNA minor groove (20). As a result, the •OH cleavage pattern is a predictor of DNA minor groove geometry and electrostatic potential (Figure 4b), as shown by PI Rohs and co-I Tullius in their first joint publication (27).

3.c.3. Experimental approach for mapping major groove shape: EPR spin labeling. Co-I Qin is leading the development and application of the site-directed spin labeling (SDSL) method for studying nucleic acids (21, 22, 40). SDSL obtains information on biomolecules (e.g., distances) by monitoring site-specifically attached stable nitroxide radicals using Electron Paramagnetic Resonance (EPR) spectroscopy. SDSL requires a small amount of sample (~50 μM in 5 μl, ~1,000 times less than NMR), and is applicable to probe biomolecules at physiological conditions without requiring a crystalline sample. Co-I Qin has developed unique SDSL methods that provide information on DNA/RNA that is difficult to obtain using other techniques (21, 22). In particular, the Qin lab developed a method for efficiently attaching a nitroxide (designated as R5) to a specific backbone phosphate group within any target DNA/RNA sequence without perturbing the native DNA/RNA structure (41, 42). Prior studies demonstrated that inter-R5 distances ranging from 20 to 50 Å can be measured using an advanced pulsed Double Electron-Electron Resonance (DEER) technique. The measured distances match those expected from known structures (42-45). In addition to using the R5 probe to map RNA conformation (46), the Qin lab has measured distances across the major groove in a number of DNA sequences (Figure 5).

3.c.3. Comparison with high-resolution structures from the Protein Data Bank (PDB). We will predict the DNA shape of all crystal structures of free and bound DNA in the Protein Data Bank that are at least one helical turn in length and have no chemical modifications. This more relevant set of unbound DNA structures is limited to only about 90 crystal structures (6). While crystallographic packing forces can affect crystal structures, solution conditions mimic the conditions within the cell. NMR structures are solved in solution, but DNA is characterized by a sparse number of NOE constraints. Therefore, only NMR data for DNA that include residual dipolar or J-coupling are sufficiently accurate for validation purposes (47), which reduces the number of available structures (6). This comparison will be based on structural parameters and RMSD and includes a collaboration with the Nucleic Acid Database (support letter from Dr. Wilma Olson).

3.c.4. Comparison with Molecular Dynamics (MD) trajectories from Ascona B-DNA Consortium (ABC).

MD simulation studies have pioneered molecular modeling approaches for unbound DNA and protein-DNA complexes (48). Much has been learned from these approaches (23, 49, 50), and MD simulations are the method of choice for studying detailed all-atom interactions between proteins, proteins and nucleic acids, and solute and solvent molecules (51-57). In the case of free DNA, the ABC consortium is playing a very important role in gathering structural information on all 136 unique tetranucleotides (23, 49, 50). These simulations are currently extended beyond the 100 ns time scale, and some of them beyond the 1 ms time scale (58). MD simulations have the
advantage of providing time scales and of accurately describing solvent interactions. While MC simulations represent a cruder sampling compared to MD studies, our main reason for proposing an MC approach is the fast conformational screening of intra-molecular interactions due to random moves and the use of a reduced set of MC variables, derived from the chemical topology of DNA and based on an implicit solvent description. In addition, the intricate balance between intra-molecular forces and interactions with the explicit solvent is difficult to capture in MD. For example, some MD simulations of nucleic acids resulted in a systematic underestimation of Helix Twist (49, 50). Following this observation, the AMBER force field (59) was significantly improved (58, 60) and led to an extensive description of B-DNA topology based on tetramers from MD trajectories (23) with which we will compare our predictions (support letter from Dr. Richard Lavery). Although this level of accuracy is not part of this proposal, we will collaborate with the ABC consortium on integrating MC and MD studies in a way that MC simulations will be used for a fast and cruder sampling of the global conformational space and MD simulations will be applied for structural refinement resulting in an accurate description of solvent-solute interactions. As part of this effort, Dr. B. Jayaram will be spending a sabbatical at USC.

3.c.5. Computational approach for high-throughput server predictions. Whereas a thoroughly validated MC simulation method is the underlying tool that we propose to further develop, the goal of this proposal is a compute server for ad-hoc genome-wide shape predictions. This server will provide all-atom models of DNA shape for DNA of any length, including whole genomes. The shape prediction server will be based on previously generated MC prediction data. Currently, we have generated approximately 2,000 MC trajectories each consisting of 150,000 all-atom structures if a snapshot of only every 10th MC cycle is taken into account. This dataset of DNA structures with energy values accepted based on the Boltzmann-Metropolis criterion (61) has mainly been generated for a study just published in Cell (19) and is therefore currently biased by an overrepresentation of Hox binding sites.

We currently use a hybrid model of a sliding-tetramer/pentamer-window approach to infer the structure of any desired DNA sequence, as illustrated in Figure 5. The conformation of the central dinucleotide in a tetranucleotide and the central nucleotide in a pentanucleotide is used to construct a predicted model of DNA. This tetramer/pentamer hybrid model takes into account nearest neighbor effects but it is an approximation of the effect of sequence on shape. Other computational studies have contributed to this approach of using tetranucleotides in inferring DNA structure (23, 62). Compared to the use of crystal structures from the PDB (62), our approach has the advantage of a complete coverage of both tetramers and pentamers. The ABC consortium uses MD simulations of sequences designed to contain all 136 unique tetranucleotides at least once (23). Compared to the pioneering effort of the ABC consortium, our MC sampling undoubtedly represents a much cruder sampling, but it has the advantage of many occurrences of each tetranucleotide in different sequence environments, and of all 512 unique pentanucleotides, even in the pool of preliminary data.

3.c.6. Aim 1. Integrate Monte Carlo simulation algorithm for DNA shape prediction with hydroxyl radical cleavage (minor groove maps) and EPR spin labeling experiments (major groove maps)

3.c.6.i. Aim 1. Preliminary data. MC simulations based on PI Rohs’ sampling method (3, 26) have already been used to study several biological questions. It was shown that high- and low-affinity binding sites of the papillomavirus E2 protein differ in their sequence-dependent bending(3). Whereas the naked high-affinity target site assumes a similar bend angle as observed in the complex with the protein, the low-affinity site is
essentially straight in its unbound form (3). These observations could be verified since high-resolution crystal structures are available for the naked sites (63, 64) and for papillomavirus E2-DNA complexes (65, 66).

The MC method was also applied to drug-DNA docking (25), and it identified minor groove shape as the origin of specificity for the Hox protein Scr (4), a finding that was recently expanded to all Drosophila Hox proteins (19). In unpublished work (Figure 7), MC simulations are used to predict the shapes of several TF-DNA binding sites. We have previously shown that minor groove geometry as observed in crystal structures of many protein-DNA complexes correlates with electrostatic potential, which in turn attracts arginines (1). This finding has had a large impact on the field.

The preliminary work for this proposal was driven by the question of whether the shape of a DNA binding site when bound to a protein is intrinsic to the sequence or induced upon protein binding. The answer to this question will improve the understanding of binding specificity on a molecular basis. Since experimental structures for the unbound forms of the sites shown in Figure 7 are not available, a computational analysis is an appropriate approach for separating intrinsic effects from effects on their conformation upon protein binding.

The structures of the protein-DNA complexes are available in the PDB, and the unbound DNA structure is predicted. MC data for naked DNA targets show that narrow regions of the minor groove are sequence-dependent features of A-tracts (67-69), which are runs of at least three consecutive ApA, ApT, or TpT base pair steps (underlined in red in Figure 7). These results demonstrate that DNA sequence induces a specific shape that leads to enhanced negative electrostatic potential, which in turn attracts positively charged arginines (1). It can now be understood how these minor groove contacts contribute to TF-DNA specificity (2).

It is important to emphasize that there is no simple one-to-one correspondence between sequence and shape. The conformation of a dinucleotide, which is described by inter-base pair helical parameters (28, 70-72), can be different depending on its adjacent base pairs, and groove geometry can even be affected by nucleotides half a helical turn away (3, 4). To test the performance of the current MC methodology, prediction results have been obtained for the Dickerson dodecamer (27), the experimentally best studied DNA molecule (6). The Drew-Dickerson dodecamer is of palindromic sequence CGCGAATTCGCG. The metric for the comparison of a prediction with a structure solved by high-resolution X-ray crystallography or NMR spectroscopy is root-mean-square-deviation (RMSD). The central hexamer of the Dickerson dodecamer (underlined), is predicted with an all-heavy atom RMSD of 1.2 Å relative to X-ray and 1.1 Å with respect to NMR structures. The MC prediction is independent of the starting configuration, as simulations started from ideal A-DNA and from ideal B-DNA, which differ by an RMSD of 5.3 Å, resulted in predictions differing by an RMSD of only 0.1 Å.

To exemplify that our MC approach correlates with protein-DNA binding properties, we studied two proteins that bind to a variety of sequences but with various binding affinity. One example is the architectural protein Factor for inversion stimulation (Fis), which binds to DNA with affinities that vary in three orders of magnitude. Crystal structures of Fis bound to a number of sequences have been reported (24), and the architecture of the complexes is very similar, which does not provide an explanation for their differential binding affinity. Therefore, we predicted the shapes of the DNA binding sites, for which Fis binding affinities are available, in their unbound state. The predictions are shown in Figure 6 for the highest ($K_d=0.2$ nM) and lowest ($K_d=140$ nM) affinity sites (solid blue and red plots represent MC predictions of free DNA; solid black plots refer to crystal structures of protein-bound DNA). As apparent from the figure, the DNA in the complex exhibits a region of narrow minor groove in its center, between the two regions contacted by the Fis dimer (24). This distinct shape is already present in the high-affinity binding site in its naked form, while the low-affinity needs to adopt this conformation upon binding. This explains why the pre-shaped site binds to the Fis dimer with much higher affinity. Thus, based on the knowledge that the shape of the five central base pairs affects binding
affinity, we performed MC simulations for all sequences for which Fis affinity data are available (24), calculated the average minor groove width for the five central base pairs and plotted this value against binding affinity.

The resulting log plot in Figure 8a demonstrates a relationship ($R^2$=0.54) between binding affinity and DNA shape at the center of the Fis binding site (red arrow in Figure 8b). Of the seven sequences, for which binding affinity was reported (24), one sequence contains a TpA base pair step in its center. The two base pairs forming a TpA step have the weakest stacking interaction among the ten unique dinucleotides (68). The so-called TpA ‘hinge’ step tends to widen the minor groove, but due to the weak stacking, it can, at low energy cost, be induced to narrow the minor groove when bound to a protein. Thus, the characteristic property of a TpA step is flexibility rather than distinct shape, and the TATA box binding protein (TBP) recognizes sequence-dependent flexibility (73, 74). Upon removal of the Fis binding site with a central TpA step, Figure 8a shows that the \textit{predictive power of binding affinity based on the shape} of naked Fis binding sites increases significantly ($R^2$=0.74).

In a different example, \textit{DNA shape is a sensitive predictor of DNase I cleavage rate.} The endonuclease DNase I interacts with double-stranded DNA through minor groove contacts. We predicted the shape of 20 hexamers in three different flanking environments and plotted minor groove width at a given position against DNase I cleavage rate (Figure 8c). These 20 sequences were selected in order to represent a wide range of cleavage rates. The resulting log plot indicates a \textit{strong relationship} ($R^2$=0.63) \textit{between DNA shape and DNase I cleavage activity}. The position in the DNA site that correlates with cleavage rate is located near the cleavage site (red arrow in Figure 8d) marking the center of the minor groove region contacted by the protein.

3.c.6.ii. Aim 1. Proposed research. In the first stage of the project, we will further \textit{develop and improve the MC methodology} based on a thorough comparison with (i) hydroxyl cleavage intensity to be measured by co-I Tullius, (ii) EPR spin labeling data to be generated by co-I Qin, (iii) high-resolution crystal and NMR structures from the PDB/NDB, and (iv) MD trajectories from the ABC consortium (23). Currently, sampling artifacts occur for very flexible sequences in less than 5% of MC simulations. The most frequent artifact is a drift towards D-

The Tullius lab is currently developing a \textit{high-throughput experimental platform, incorporating laboratory robotics and capillary electrophoresis}, to greatly increase the number of sequences for which hydroxyl cleavage intensity can be measured. These enhancements in experimental technology will enable a close interaction between computational shape prediction and experimental validation. Co-I Tullius will measure the •OH cleavage patterns of sequences for which all-atom structures have been predicted by PI Rohs using his MC approach. To do this, co-I Tullius will use his high-throughput robotics/capillary electrophoresis platform to measure the hydroxyl radical cleavage patterns of specific DNA sequences. We will \textit{modify the force field} to remove this artifact \textit{by implementing weak constraints for minor and major groove width}, as illustrated in Figure 9, which will be derived from experimental data to be generated by co-I Tullius (minor groove map) and co-I Qin (major groove map).

The compelling metric of relating •OH cleavage intensity to a \textit{surface area} (Figure 4), has been \textit{expanded by Rohs and Tullius to a metric that relates •OH cleavage intensity to the accessible surface area (SASA)} is directly proportional to measured •OH cleavage intensity due to the differential accessibility of sugar hydrogen atoms. This \textit{compelling metric of relating •OH cleavage intensity to a surface area} (Figure 4), has been \textit{expanded by Rohs and Tullius to a metric that relates •OH cleavage intensity to the accessible surface area (SASA)
**intensity to minor groove width** (27) yielding a direct measure of distance. As demonstrated in this very recent publication, •OH cleavage intensity provides a map of minor groove shape (27). We are now using this metric to compare and integrate MC prediction with •OH cleavage experiments by defining distances based on the relationship shown in Figure 4b.

While •OH cleavage data for all 136 unique tetranucleotides are already available in the ORChID database (20), co-I Tullius will now measure •OH cleavage for about 100 TF binding sites, such as the ones shown in Figure 7. After the first year, we plan to generate experimental and computational data on longer DNA binding sites including the following examples: (i) the glucocorticoid receptor (75), which binds to half sites separated by spacers of three base pairs, and other nuclear receptors, (ii) the tumor suppressor p53 (5, 76, 77), which binds to half sites separated by spacers of variable length, and (iii) the IFN-β enhanceosome (78-80), on which five TFs assemble at adjacent sites.

This information will refine the ORChID database (20, 27) and provide valuable information on the effect of nearest neighbors of tetramers on minor groove geometry. Based on the metric introduced here, **we will define harmonic functions that represent weak constraints for minor groove width**, shown in Figure 9a. These harmonic terms will be added to the force field and make the MC sampling faster, more efficient, and prevent deformations and deviations from B-DNA. Since minor and major groove width are only loosely anti-correlated ($R^2 = –0.12$ for the current MC dataset based on CURVES), we chose to propose an approach complementary to •OH cleavage. We will use the EPR site-directed spin labeling (SDSL) technique (21, 22) to measure distances across the major groove. Based on these distances, we will also derive weak **constraints for major groove width**, shown in Figure 9b, and **add related harmonic terms to the force field, which concludes the feedback loop**.

For this proposal, the Qin lab will use the established R5 tool-kit to measure distances across the major groove in DNA duplexes, as illustrated in Figure 5a, containing all 136 unique tetranucleotides. As noted, preparation of R5-labeled DNAs is highly efficient, both in terms of time and cost (42). Furthermore, an **NIH-funded state-of-the-art pulsed EPR spectrometer** (NCRR 1S10RR028992-01; PI: Qin) **has been installed at USC and is now fully operational**, thus providing ample spectrometer time for the proposed measurements. Based on preliminary studies, we expect that 5-8 DNA datasets will be collected every month. Therefore, the proposed project duration is sufficient for collecting all 136 data sets, including repeat measurements. Upon completing work on the tetranucleotides, measurements will be expanded to the pentanucleotide units.

**This integration of MC sampling with •OH cleavage and SDSL experiments will form a feedback loop.** First we will perform MC simulations for a large number of sequences. In the second step, we will compare our MC predictions with the experimental data. In a third step we will derive distance metrics across the minor and major groove from •OH cleavage and SDSL data (Figure 9) and use those distances to refine the MC sampling. We will add harmonic terms to the force field, which represent the measured distances as weak constraints that allow a flexibility of the measured distances of about ±1-2 Å (exact value to be refined) while guiding the conformational sampling. The goal of this proposal is to achieve wide sequence coverage, in terms of all 2080 unique hexamers in many different flanking sequence environments. MD simulations are of course much more accurate on an all-atom basis, and describe solute-solvent interactions explicitly, but we propose using the cruder sampling in MC simulations, guided by additional constraints derived from experiments, for generating the underlying data for high-throughput DNA shape predictions on a genomic scale.

The current force field that is being used is the Cornell et al. AMBER force field (59), which is widely tested for nucleic acids (48), with a slight modification disfavoring BII over BI conformations. We will **add a new version of AMBER** (58), and also **CHARMM** (81) **and other force fields** in order to allow for other developers to implement our approach. We will also add the possibility to **increase the salt concentration** and work with **multivalent ions**. We will add an option to work with **ligands** that can be moved relative to the DNA and varied based on their internal degrees of freedom (25). And we will offer an option to replace the distant-dependent dielectric solvent model (82) by the **Generalized Born solvent model** (83). We will also add an option for replacing the current deoxyribose sampling in phase and amplitude by the four torsion angles describing the sugar puckering. This will increase the number of MC variables per nucleotide from 12 to 14 but remove the approximate pseudo-rotation description of the deoxyribose ring (26). Finally, we will further
increase the sampling efficiency by implementing a recently published chain closure algorithm that accepts an approximate chain closure in cases when a perfect closure is not possible (84).

These methodological developments are substantive changes to the current software and will, therefore, require extensive testing and validation. We will use high-resolution X-ray and NMR structures from the PDB for validation purposes. As a high-throughput approach for inferring DNA structure in solution, we will use hydroxyl cleavage intensity measurements performed by co-I Tullius. Validation based on •OH data will focus on minor groove shape, which plays a key role in achieving protein-DNA binding specificity (1, 4, 19). Although major groove shape was not observed yet to determine binding specificity due to the dominant role of base specific hydrogen bonds in the major groove, we will map the major groove geometry for the 136 unique tetranucleotides by EPR spin labeling experiments performed by co-I Qin. The combination of MC sampling with two experimental approaches is expected to provide a basis for large-scale prediction. Our goal is to improve the software that implements our MC method so that we can publicly release the source code to biomedical researchers within the first two years of the project. We will then develop a wiki page where we will receive and respond to feedback, and which we expect to lead to further improvements.

3.c.7. Aim 2. Develop compute server for high-throughput DNA shape prediction on genome-wide basis

3.c.7.i. Aim 2. Preliminary data. The existing MC trajectories generated in previous MC simulations will be used to infer DNA shape in ad-hoc high-throughput predictions. For this purpose, we adopted a sliding-k-mer-window approach, and our first step was to derive the conformation of all possible tetranucleotides. The four nucleotides A, C, G, and T form 136 unique tetranucleotides, considering that the two DNA strands of a duplex are identical, thus AAAA is equal to TTTT. Based on this approach, our current MC data leads to coverage that includes 100% of the 136 unique tetranucleotides. In comparison, the coverage of crystal structures of free DNA deposited in the PDB only comprises 95 of 136 unique tetranucleotides. The insufficient coverage based on experimental structures emphasizes the need to close this gap with computational methods (6).

The tetramer model used for the ad-hoc compute server prediction of DNA shape is explained in Figure 6a-d. We applied this approach in preliminary tests for predictions of high- and low-affinity binding sites of the architectural Fis protein, for which crystal structures have been recently published (24), and realized that using a tetramer model is not sufficient. This advancement is illustrated in Figure 6e-f where we use a hybrid model for predicting the shape of Fis binding sites. There are 512 unique pentanucleotides and we just reached 100% pentamer coverage based on current MC data. Due to the lower frequency in occurrence, we used a hybrid model in Figure 6e-f, which bases its prediction on pentamers above a certain threshold of occurrence, and on tetramers below that threshold. The advantage of using pentamers is apparent in Figure 6.

3.c.7.ii. Aim 2. Proposed research. Although the work presented here is preliminary, the difference in shape that is seen in MC simulations can already be inferred based on the proposed hybrid model of tetra- and pentanucleotides. At the current stage, we only derive minor groove width based on the central di-nucleotide of tetramers combined with the central nucleotide of pentamers. This model has already been applied on thousands of Hox binding sites in a recent publication in Cell (19). We are currently generating more MC data in order to increase our pentamer coverage depth. One goal for Aim 2 is to develop a server based on a broad coverage of hexamers. Currently, there are about 64% of all 2,080 unique hexamers present in the dataset of MC trajectories. Another goal is to calculate all structural parameters that characterize DNA and to derive all atom-models. We will use 3DNA (72, 85) for building all-atom structures based on the conformation of the central dinucleotide of a hexamer. Such all-atom models can then be analyzed in detail in structural and energetic terms, based on 3DNA (72, 85), CURVES (28, 86, 87) and DelPhi (29, 30, 88, 89) interfaces. In addition, the MC simulation data used for the current hybrid model will be updated and significantly extended to increase sequence coverage. Therefore, MC data generation will be a major ongoing activity.

The sliding window approach described above is somewhat similar to the approach used in the ORChID server (20, 36), which infers cleavage intensities based on a sliding tetramer window algorithm for sequences that have not been studied experimentally. The similarity of the approach allows for a direct comparison. We will combine both tools in a joint server. Co-I Tullius’ current ORChID server (36) has been used to predict the DNA structural profile of the entire human genome. These structural data have been deposited in the UC Santa
Cruz Genome Browser and are publicly available. Our goal is to allow access to our combined new server via the same gateways. ORChID provides information on the SASA of deoxyribose groups while our proposed DNA shape prediction server generates all-atom models, which can be analyzed for a wide array of structural information (e.g., major and minor groove shape, DNA bending and torsions, flexibility).

3.c.8. Aim 3. Apply high-throughput DNA shape prediction to genome-wide studies of transcription factor-DNA binding, and derive sequence-shape logos and shape heat maps

3.c.8.i. Aim 3. Preliminary data. DNA shape analysis becomes valuable if large amounts of genomic data can easily be analyzed. The question then arises how shape can be represented and interpreted. Figure 11 shows the shape analysis of in total > 650,000 binding sites of the 8 Drosophila Hox proteins, including two isoforms of Ubx, in complex with their cofactor Exd, which Rohs just published in Cell (19). Hox proteins need to recognize subtle differences in their in vivo sites with high specificity in order to execute their distinct functions in embryonic development (90). Since the nucleotide sequences of the in vivo sites of Hox proteins are similar, our hypothesis was that shape is an additional discriminator in order to achieve a high degree of binding specificity. Our most recent work confirms this hypothesis, and moreover, indicates that Hox genes differentiated throughout evolution based on DNA shape (19). Our collaborators used SELEX combined with deep sequencing (91-94), to identify a set of DNA sequences selected by Drosophila Hox proteins. The SELEX-seq experiments mimic the biological assembly due to the presence of both cofactors. As a result, the data reflect a higher degree of specificity than previously known for monomeric homeodomain binding (95, 96).

Minor and major groove profiles were subsequently predicted based on the preliminary version of our high-throughput server, for all SELEX-seq reads with a relative affinity above 0.1 with respect to the top binder in each category. Figure 11 compares the results for the DNA binding sites selected by the 8 Hox proteins. The data shows that the minor groove geometry is distinct for binding sites of anterior vs. posterior Hox proteins (Figure 11a-b) whereas the major groove geometry is not expected to contribute to differential binding specificity. In the minor groove, the main difference occurs at the second AY position of the general Exd-Hox motif AYNAY (Y=C or T). The predicted distinct minor groove shape was also observed in crystal structures of the Hox proteins Scr and Ubx in ternary complexes with their cofactor Exd and DNA. This observation is, therefore, another validation of the proposed shape prediction techniques.

3.c.8.ii. Aim 3. Proposed research. To establish shape as an important property of DNA, we will curate heat maps representing major and minor groove shape for all known TF binding sites. In addition, we will develop novel logo representations that contain shape information, to convey shape information in a similar way as sequence in WebLogos (7, 97). Examples of sequence-shape logos, currently based on minor groove width due to its larger role in TF specificity, are shown in Figure 12. For the purpose of combining sequence and shape, we combined the four letters A, C, G, and T from standard sequence logos with two fonts representing shape. We use the empty letter font for indicating regions that exhibit a narrow minor groove of < 5 Å, and the regular font for representing regions with a medium to wide minor groove > 5 Å. This simplest version of sequence-shape logos is shown in Figure 12. Whereas each position in a sequence logo is independent of its neighbors, a sequence-shape logo incorporates interrelationships between positions because shape is a result of the overall sequence composition. The
information content of a logo increases with the number of sites taken into account. We will apply this concept to the analysis of sequence information, such as ChIP-seq data (98-100). The innovative potential of this approach is that a combination of sequence and shape can now be used in a motif search for TF binding sites (8, 101). This approach has been tested for a number of datasets. As shown in Figure 12a-b for the analysis of ChIP-seq data for the CCCTC binding factor (CTCF), the motif finding algorithm MEME (8) identifies the correct motif, both when only sequence or a combination of sequence and shape information is used. ChIP-seq data for CTCF is known to be of high quality, and this data is therefore used to validate the method, and as shown motif finding based on sequence or sequence-shape logos identifies the same TF binding motif. In a different example of an Arabidopsis homeobox TF, it was not possible to find a motif in ChIP-seq data based on sequence alone. As shown in Figure 12c, only when a combination of sequence and shape information is used, the common TAAT homeobox motif can be identified, using MEME (8). This discovery of an otherwise undetected motif illustrates the significance of this proposal. We will develop the methodology for (i) inferring shape heat maps and sequence-shape logos and (ii) for using them in motif search. We will make this tool accessible, and we will build a public database in which we will curate sequence logos, shape heat maps, and sequence-shape logos for all known TF bindings sites.

3.c.9. Aim 4. Study the effect of DNA shape on nucleosome formation and chromatin organization.

3.c.9.i. Aim 4. Preliminary data. The histone octamer binds largely non-specifically to DNA. When DNA wraps around the histone core, the minor groove faces the histones every ten base pairs. Periodic distributions of dinucleotides have been associated with nucleosome positioning signals (9, 13, 14). PI Rohs and collaborators stressed that short A-tracts appear with a periodicity of a helical turn (1, 102). The ability to predict DNA shape in a high-throughput manner, will allow us to go a step further to see if DNA structure shows the same periodicity that is observed in nucleotide sequence for nucleosome binding sites. Since each groove of DNA when bound in nucleosomes faces the histone core every 10 base pairs, the periodicity of a helical turn assists in generating the deformation necessary for forming nucleosomes. In preliminary work, we predicted the DNA shape of 23,076 sequences from the yeast genome and 25,654 sequences from the fly genome, which were experimentally found occupied by nucleosomes in vivo. Figure 13 clearly indicates a periodicity of ten base pairs in average minor groove width of these sequences, even without histones present. This finding is congruent with an analysis based the ORChID2 server, which Tullius and Rohs recently published (27).

3.c.9.ii. Aim 4. Proposed research. The first question that we will study is how the different sequence composition in various genomes affects DNA shape and its impact on nucleosome formation. Of particular interest is a comparison of nucleosome occupied regions in the yeast (9, 13, 14), Drosophila (15), and E. coli genome (10) with the genome of the malaria parasite Plasmodium falciparum, which has an extremely high A/T content of about 80% (16, 103). Is there a periodicity in ApT vs. TpA stacking interactions in a genomic alphabet, which has been reduced to (almost) two letters? Is DNA shape being used to encode function? In order to answer these questions, we will compare (i) the sequence composition (e.g., A-tracts vs. TpA steps) and (ii) predict structure based on our high-throughput compute server. Another question that we propose to ask is whether there is a different functional role of A/T vs. G/C regions. It has previously been proposed that an increased G/C content is a nucleosome positioning signal (104). Is the function of the G/C regions the separation of A/T regions? High-throughput shape predictions in comparison with sequence analysis are expected to answer this and similar questions.

3.c.10. Summary. We propose an analysis of the role of DNA shape and flexibility, at whole genome lengths, on TF-DNA binding specificity, nucleosome formation, and genome organization. Our new tools will transform DNA shape analysis into a high-throughput approach for analyzing genome-wide sequence information. Our findings will advance our understanding of the role of intrinsic DNA shape and flexibility, and its sequence-dependent response to proteins in fulfilling a wide range of unique biological functions.


47. Wu Z, Delaglio F, Tjandra N, Zhurkin VB, Bax A. Overall structure and sugar dynamics of a DNA
dodecamer from homo- and heteronuclear dipolar couplings and 31P chemical shift anisotropy. Journal
Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides.
50. Dixit SB, Beveridge DL, Case DA, Cheatham TE, 3rd, Giudice E, Lankas F, Lavery R, Maddocks JH,
Osman R, Sklenar H, Thayer KM, Varnai P. Molecular dynamics simulations of the 136 unique
tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical
53. Pan Y, Nussinov R. p53-Induced DNA bending: the interplay between p53-DNA and p53-p53
54. Pan Y, Nussinov R. Cooperativity dominates the genomic organization of p53-response elements:
55. Pan Y, Nussinov R. Preferred drifting along the DNA major groove and cooperative anchoring of the
56. Pan Y, Nussinov R. Lysine120 interactions with p53 response elements can allosterically direct p53
57. Pan Y, Nussinov R. The role of response elements organization in transcription factor selectivity: the
58. Perez A, Marchan I, Svozil D, Sponer J, Cheatham TE, 3rd, Laughton CA, Orozco M. Refinement of the
AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. Biophys J.
59. Cornell WD, Cieplak P, Bayly CI, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell
JW, Kollman PA. A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and
61. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by
63. Rozenberg H, Rabinovich D, Frolow F, Hegde RS, Shakked Z. Structural code for DNA recognition
revealed in crystal structures of papillomavirus E2-DNA targets. Proc Natl Acad Sci U S A.
64. Hegde RS, Grossman SR, Laimins LA, Sigler PB. Crystal structure at 1.7 A of the bovine
65. Kim SS, Tam JK, Wang AF, Hegde RS. The structural basis of DNA target discrimination by


