Toward an Expanded Genome: Structural and Computational Characterization of an Artificially Expanded Genetic Information System

Nigel G. J. Richards†,‡ and Millie M. Georgiadis*,§,∥

†School of Chemistry, Cardiff University, Cardiff CF10 3AT, United Kingdom
‡Foundation for Applied Molecular Evolution, 13709 Progress Boulevard, Alachua, Florida 32615, United States
§Department of Biochemistry & Molecular Biology, Indiana University School of Medicine, Indianapolis, Indiana 46202, United States
∥Department of Chemistry and Chemical Biology, Indiana University–Purdue University Indianapolis, Indianapolis, Indiana 46202, United States

CONSPECTUS: Although the fundamental properties of DNA as first proposed by Watson and Crick in 1953 provided a basic understanding of how duplex DNA was organized and might be replicated, it was not until the first crystal structures of DNA (Z-DNA in 1979, B-DNA in 1980, and A-DNA in 1982) that the true complexity of the molecule began to be appreciated. Many crystal structures of oligonucleotides have since shed light on the helical forms that “Watson–Crick” DNA can adopt, their associated groove widths, and the properties of the nucleobase pairs and their interactions in all three helical forms. Additional understanding of the properties of Watson–Crick DNA has been provided by computational studies employing a variety of theoretical methods. Together with these studies devoted to understanding Watson–Crick DNA, recent efforts to expand the genetic alphabet have founded a new field in synthetic biology. One of these efforts, the artificially expanded genetic information system (AEGIS) developed by Steven Benner and co-workers, takes advantage of orthogonal hydrogen bonding to produce DNA comprised of six nucleobase pairs, of which the most extensively studied is referred to as P:Z with P being 2-amino-imidazo[1,2-a]-1,3,5-triazin-4(8H)-one and Z being 6-amino-5-nitro-2(1H)-pyridone. P:Z forms three edge-on hydrogen bonds that differ from standard Watson–Crick pairs in the arrangement of acceptors and donor groups; P presents acceptor, acceptor, donor, and Z presents donor, donor, acceptor. Z is unique among the AEGIS nucleobases in having a nitro group present in the major groove. PZ-containing DNA has been exploited in a number of clinical applications and is being used to develop receptors and catalysts. Ultimately, the grand challenge will be to create a semisynthetic organism with an expanded genome. Furthermore, just as our understanding of the properties of natural DNA have benefited from structural and computational characterization, so too will our understanding of artificial DNA. This Account focuses on the structural and biophysical properties of AEGIS DNA containing P:Z pairs. We begin with the fundamental properties of P:Z nucleobase pairs, including their electrostatic potential and hydrogen-bonding energies, as elucidated by quantum mechanical calculations. We then examine the impact of including multiple consecutive P:Z pairs into duplex DNA providing an opportunity to investigate stacking interactions between P:Z pairs. The self-complementary 5′-CTTATPPTAZZATAAG was crystallized in B-form using the host–guest system along with analogous natural sequences including Gs or As. Use of the host–guest system to characterize B-DNA obviates a number of limitations on the structural characterization of sequences of interest; these include the ability to crystallize the desired sequences and to distinguish structural effects imparted by the lattice constraints from those inherent in the sequence itself. On the other hand, 3′/6ZP, 5′-CTTATPPZZZZATAAG, was crystallized in A-form in a DNA-only lattice allowing a comparative analysis of P:Z pairs in two of the biologically relevant helical forms: A- and B-DNA. Computational studies on the 3′/6ZP sequence starting in A-form provide additional evidence for a more energetically favorable stacking interaction, which we term the “slide” conformer, observed in the A-form crystal structure; this unusual stacking interaction plays a major role in altering the conformational dynamics observed for the PZ-containing duplex as compared to a GC-containing “control” duplex in long time scale molecular continued...
ARTIFICIAL DNA

One of the most important outcomes of modern synthetic biology is the recognition that the biopolymers resulting from four billion years of biological evolution are not the only molecules that might support genetics, inheritance, evolution, and catalysis. As a consequence of efforts to create artificial DNA and thereby an expanded genome capable of Darwinian evolution, we now have a much greater understanding of the fundamental properties of natural DNA. Although the phosphodiester backbone has been shown to be essential for DNA structure and function, the nucleobases can and have been broadly altered, resulting in pairs that rely on orthogonal hydrogen bonding patterns or those that dispense with hydrogen bonding and rely on steric and hydrophobic complementarity. The most extensive and widely used of these is an artificially expanded genetic information system (AEGIS) that employs orthogonal hydrogen bonding patterns between big purine-like and small pyrimidine-like nucleobases maintaining a duplex DNA structure. Thus, the number of nucleobase pairs can be increased from two to six by merely rearranging the pattern of hydrogen bond donor and acceptor groups.

Practical efforts to implement this idea using chemical synthesis have thus far led to several "generations" of novel heterocycles that can be employed in automated DNA synthesis, thereby yielding artificially expanded alphabets. DNA molecules containing the P:Z nucleobase pair (Figure 1) have proven of special interest given that several polymerases will replicate them in nested PCR reactions, and Taq DNA polymerase variants have been obtained that will efficiently and faithfully replicate this base pair. A unique feature of Z is the nitro group in the major groove providing additional functionality as compared to a natural nucleobase. P:Z and possible mismatches have been extensively studied through UV absorbance melting measurements and determination of the energetics of binding of DNA strands, providing evidence that P:Z pairs contribute more to duplex stability than any mismatches involving either nucleobase with natural nucleobases and that P:Z is more stable than the most similar natural pair G:C. In addition, protein engineering using a variety of strategies has created polymerases that copy and amplify oligonucleotides containing another AEGIS pair: X and K (Figure 1).

On the other hand, exploiting altered patterns of hydrogen bonding to obtain novel nucleobase pairs that are "orthogonal" to A:T and G:C has proven surprisingly problematic. For example, many heterocycles have populated tautomeric forms with altered hydrogen bonding patterns; these can base pair with standard nucleobases in either duplex DNA or within the active sites of polymerases, thereby giving rise to unanticipated mutations or the loss of the AEGIS nucleobases during replication. Even if these "design" problems were to be easily resolved, little is known about how the incorporation of these non-natural nucleobases into DNA affects the conformational preferences and dynamical properties of this complex molecule, which is fundamental to the interaction of "standard" DNA with proteins, such as polymerases and transcription factors. Indeed, the first studies aimed at understanding how P:Z nucleobase pairs, which have altered electrostatic properties (dipole moments, charge distribution), might perturb DNA structure and dynamics have only recently appeared. Moreover, molecular insights into how proteins might recognize AEGIS DNA molecules have not yet been reported, and we note that predictive computational assessments of the interaction free energies between AEGIS-based DNA and proteins will considerably aid efforts to create reagents for using AEGIS molecular components in bacterial cells with all of the associated implications for synthetic biology. Although many challenges remain, non-natural nucleo-base pairs have now been successfully replicated in a bacterial system, paving the way for the development of semisynthetic organisms.

QUANTUM MECHANICAL CHARACTERIZATION OF THE P:Z PAIR

As a prelude to examining the global impact of multiple consecutive P:Z nucleobase pairs on the DNA duplex, we examined the electrostatic potential and hydrogen bonding properties of the two non-natural nucleobases. Considerable precedent exists in using quantum mechanical (QM) calculations for modeling the electronic properties of individual Watson-Crick nucleo-bases and their associated base pairs, especially in combination with explicit and continuum solvation models. For example, such computational studies have established the energies of A:T and G:C hydrogen bonding, both in vacuo and in aqueous solution, and rationalized the energetic preference for unsymmetrical hydrogen bond arrangements as in A:T and G:C rather than the symmetrical pattern that is present in X:K nucleobase pairs. To date, QM calculations have only been reported for the P:Z nucleobase pair. Although high-level ab initio calculations suggest that the free energy of hydro-gen bonding in the P:Z nucleobase pair is less favorable (1.4 kcal/mol) than for G:C in the gas-phase, calculations that include the effects of solvation indicate that the presence of the Z-nitro group decreases the enthalpy of the hydrogen bonds in P:Z relative to those in G:C by 0.6 kcal/mol. This finding is supported by recent biophysical studies in which the P:Z hydrogen bonding interactions were found to be stronger than natural or mispaired interactions involving either P or Z.

Figure 1. Chemical structures of P:Z, X:K, G:C, and A:T nucleobase pairs. R is 2'-deoxyribose in duplex DNA.
Little systematic computational work has been performed to examine how placing AEGIS nucleobases within the DNA duplex modifies their electronic distributions, but simple gas-phase QM calculations on isolated P:Z and G:C nucleobase pairs clearly show significant differences in dipole moments and electrostatic potential within both grooves (Figure 2), which can potentially be exploited in re-engineering the specificity of transcription factors and restriction endonucleases.\(^{22}\)

### Structural Characterization of P:Z Pairs in B-DNA

#### General Considerations

Crystallization of a polyanionic molecule like DNA poses a significant challenge due to the limited number of sites available for intermolecular contacts that are required to form a three-dimensional lattice.\(^{23}\) Thus, the ability to crystallize DNA alone and specifically in B-form has in the past been limited to specific sequences and lengths of oligonucleotides with the first example of a B-form DNA structure being the Drew–Dickerson dodecamer.\(^{24}\) The majority of structural studies on B-form DNA have been performed on oligonucleotides that are 12 base pairs or shorter in length. The problem of crystallizing DNA is compounded by inclusion of non-natural nucleobases, which represent uncharted territory for structural analyses. Some of these problems can be circumvented through the use of a host–guest system developed by Georgiadis and co-workers in which the N-terminal fragment of Moloney murine leukemia virus reverse transcriptase (MMLV RT) serves as the host and a self-complementary 16-mer oligonucleotide as the guest.\(^{25}\) In the complex, one DNA duplex is bound to two protein molecules; protein–DNA interactions involve the terminal three nucleo-base pairs with R116 bound in the minor groove and other interactions involving backbone atoms (Figure 3). To date, this system has been used to determine 24 natural and two artificial DNA oligonucleotide structures including some with ligands bound to the DNA (Table 1). We recognized that this type of complex could in fact serve as a host–guest system; the binding site for DNA within the protein is general and could potentially accommodate any 16-mer DNA sequence. The unique repeating unit or the asymmetric unit of the crystal includes only one protein molecule and half of the DNA molecule (Figure 3). Thus, the system is best suited to self-complementary 16-mer DNA oligonucleotides but has been successfully used to analyze the structure of sequences that are not self-complementary.\(^{26}\)

---

**Figure 2.** Dipole moments (arrows) and electrostatic potential (rendered on the VDW surface) for P:Z (left) and G:C (right) nucleobase pairs. Electrostatic energies range from approximately \(-40\) kcal/mol (red) to approximately \(+40\) kcal/mol (blue). Taken from ref 19 and used with permission.

**Figure 3.** Arrows from the duplex DNA indicate the positions of the P:Z pairs and end in a stick-rendering of the P:Z pair, top of figure. An arrow from the protein-DNA interface in the complex points to a close-up rendering of this interaction mediated by binding of R116, positioned through hydrogen bonding interactions with D114, to O2 of the second base from the end in the minor groove, shown at the bottom of figure.

The utility of the host–guest system extends to the crystallization and analysis of DNA sequences of interest\(^{25,27}\) as well as DNA–ligand complexes.\(^{28}\) The host–guest system has three major advantages over DNA-only systems: it allow us to (i) crystallize any 16-mer DNA oligonucleotide sequence that adopts B-form, (ii) phase the structure of the complex using the host as a search model in molecular replacement calculations providing unbiased electron density for the DNA, and (iii) analyze DNA structures, typically determined at 1.7–1.8 Å, that have been obtained in the same lattice and are therefore subject to the same constraints allowing DNA sequence-specific features to emerge. Crystals of desired DNA sequences are grown in a low salt, PEG 4000-containing precipitant and can be obtained rapidly through microseeding with seeds created from crystals of a standard host–guest complex.\(^{27,29}\) The major limitation of the system is that 16-mer oligonucleotide duplexes must be used to obtain crystals and that the 16-mer oligonucleotide must exist predominantly as B-form DNA.

**Structural Properties of 5′-CTTATPPTAZZATAAG**

The host–guest system is ideally suited to the rapid analysis of artificial DNA. Thus, the most recent application of the host–guest system has been the crystallization and analysis of artificial DNA including P:Z nucleobase pairs (Figure 3).\(^{30}\) Although we could easily have generated structures including a single P:Z nucleobase pair within a natural DNA environment, we were much more interested in the effects of including consecutive P:Z nucleobase pairs on the stacking interactions.
Table 1. DNA Sequences Analyzed Using the Host–Guest System

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>type&lt;sup&gt;a&lt;/sup&gt;</th>
<th>DNA sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>4XN0</td>
<td>AEGIS 3′/6ZP</td>
<td>5′-CTTTPPPPZTTAAAG&lt;sup&gt;30&lt;/sup&gt;</td>
</tr>
<tr>
<td>4XO0</td>
<td>AEGIS 2P</td>
<td>5′-CTTATPPTAZZATAAG&lt;sup&gt;30&lt;/sup&gt;</td>
</tr>
<tr>
<td>4XPC</td>
<td></td>
<td>5′-CTTATAATTATTAAAG&lt;sup&gt;30&lt;/sup&gt;</td>
</tr>
<tr>
<td>4XPE</td>
<td></td>
<td>5′-CTTATGGCCCTAAAG&lt;sup&gt;30&lt;/sup&gt;</td>
</tr>
<tr>
<td>4M94</td>
<td>spore product</td>
<td>5′-ATCCGtATAACGGATG&lt;sup&gt;31&lt;/sup&gt;</td>
</tr>
<tr>
<td>4M95</td>
<td></td>
<td>5′-ATCCGTATATCCGGAT&lt;sup&gt;31&lt;/sup&gt;</td>
</tr>
<tr>
<td>2KR</td>
<td></td>
<td>5′-ATTAGTTAATACTAA&lt;sup&gt;32&lt;/sup&gt;</td>
</tr>
<tr>
<td>2KS</td>
<td></td>
<td>5′-ATTAGTTAATACTAA&lt;sup&gt;32&lt;/sup&gt;</td>
</tr>
<tr>
<td>3FSI</td>
<td>MG lig 4,4′bP</td>
<td>5′-CTTAAITCGAAATTAAG&lt;sup&gt;33&lt;/sup&gt;</td>
</tr>
<tr>
<td>2FJv</td>
<td>MG lig RT29</td>
<td>5′-CTTAAITCGAAATTAAG&lt;sup&gt;33&lt;/sup&gt;</td>
</tr>
<tr>
<td>2FW</td>
<td></td>
<td>5′-CTTAAITCGAAATTAAG&lt;sup&gt;33&lt;/sup&gt;</td>
</tr>
<tr>
<td>2FX</td>
<td></td>
<td>5′-CTTAAITCGAAATTAAG&lt;sup&gt;33&lt;/sup&gt;</td>
</tr>
<tr>
<td>1Z7T</td>
<td>MG lig netropsin</td>
<td>5′-CTTAAITCGAAATTAAG&lt;sup&gt;33&lt;/sup&gt;</td>
</tr>
<tr>
<td>1N4L</td>
<td>HIV PPT</td>
<td>5′-CTTTTAAAAAGAAAA&lt;sup&gt;26&lt;/sup&gt;</td>
</tr>
<tr>
<td>2FVP</td>
<td>LTR</td>
<td>5′-TTTCTTCAATGAAA&lt;sup&gt;27&lt;/sup&gt;</td>
</tr>
<tr>
<td>2FVQ</td>
<td>LTR</td>
<td>5′-TTTCTTCAATGAAA&lt;sup&gt;27&lt;/sup&gt;</td>
</tr>
<tr>
<td>2FVR</td>
<td>LTR</td>
<td>5′-TTTCTTCAATGAAA&lt;sup&gt;27&lt;/sup&gt;</td>
</tr>
<tr>
<td>2FVS</td>
<td>LTR</td>
<td>5′-TTTCTTCAATGAAA&lt;sup&gt;27&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup>It refers to the spore product thymine dimer; MG lig, minor groove binding ligand; bleo B2, bleomycin B2; bIP, 4,4′-bis(imidazolyl-lamino) diphenylamine; PPT, polypurine tract; LTR, long terminal repeat. The HIV PPT sequence is shown for one of the two strands that make up this duplex.

and overall structural properties of the DNA to determine whether they had special properties. A-tracts are known to confer rigidity to the structure of DNA, whereas having multiple Gs spaced appropriately enables the formation of quadruplex structures. Thus, we pursued the structural characterization of self-complementary 16-mer sequences including either two consecutive P:Z pairs (2P, 5′-CTTATPPTAZZATAAG) or six consecutive P:Z pairs (3/6ZP, 5′-CTTATPPPPPZTTAAAG). The 2P sequence crystallized in the host-guest system with B-form. The 3/6ZP oligonucleotide did not, suggesting that another useful property of the host-guest system is to discriminate between sequences that stably adopt B-form and those that exist in solution as a mixture of different helical forms. This conclusion is supported by the molecular dynamics studies discussed below. We also found this to be true for X:K nucleobase pairs.

A 16-mer including two consecutive X:K nucleobase pairs crystallized readily in the host-guest system, whereas one including six consecutive X:K pairs did not. Oligonucleotides including 5′-GGGGCC or 5′-AAATTT sequences replacing the ZP-containing regions readily crystallized in the host-guest system (Table 1). Analysis of all of these oligonucleotides in low salt conditions by CD suggested that the P:Z sequences adopt structures that resemble those of the G:C control oligonucleo-tide. The 2P to date is the only structure of P:Z pairs in B-form DNA. A crystal structure of a 9-mer oligonucleotide, 5′-G 5-MeSedUGT-Z-ACAC-3′ and complement 5′- G 5-MeSe-dUGT-P-ACAC-3′, including Se-modified nucleotides, crystallized with four molecules in the asymmetric unit, which are either partially or fully A-form. In this case, there are no reference structures including G:C or A:T crystallized in the same lattice available for comparison; thus, it is difficult to draw any specific conclusions regarding the impact of including P:Z in this sequence.

Individual P:Z base pairs at positions 6 and 7 in the oligonucleotide sequence (see Figure 3 for numbering scheme) more closely resemble G:C than A:T pairs located in the same position in the control host–guest complexes with similar shear, stretch, stagger, and propeller values as assessed by 3DNA (Figure 4). The P:Z pair differs from both G:C and A:T pairs at position 6 in that its buckle angle is ~11.9° as compared to ~5.6° for G:C and ~1.4° for A:T at the same position (Figure 4) and presents a unique pattern of electronegative atoms in the major groove (Figure 5). Of particular interest is the zwitterionic nitro group at position 5 of the pyrimidine-like heterocyclic ring (Figure 1), which provides additional functionality in the major groove. Within the minor groove, P:Z presents hydrogen bond acceptors O2 from the pyrimidine-like Z and N3 from the purine-like P as found in all natural base pairs (Figures 1 and 5). The major groove width associated with the two P:Z pairs (18.7 Å) is on average 0.7 Å wider than that for the G:C pairs and 0.4 Å narrower than that observed for A:T pairs at the same positions as calculated in 3DNA. The minor groove width for the P:Z pairs (12.5 Å) is very similar on average to that observed for G:C (12.4 Å) and much wider than that observed for A:T (9.7 Å).

![Figure 4. P:Z in B-form DNA is sheared by ~0.87 Å and G:C by ~1.40 Å, whereas the equivalent A:T pair does not exhibit significant shearing (~0.22 Å) as shown on the right-hand side of the figure with stick renderings and filled rings for each of the nucleobase pairs. Shearing defines displacement along the hydrogen-bonding edge of one base with respect to the other. Although sheared, both P:Z and G:C retain standard hydrogen bonding distances. On the left are shown the same base pairs in an edge-on view. In this view, it is evident that buckle angle of ~11.9° for P:Z is much larger than that in G:C or A:T of ~5.6° and ~1.4°, respectively. The buckle angle defines the degree of nonplanarity across the base pair.](image-url)
Van der Waal renderings of the major and minor groove faces of the P:Z, G:C, and A:T base pairs are shown with N atoms in blue, O in red, P in orange, and C in yellow for P:Z, green for G:C, and pink for A:T.

Crystal structure of P:Z pairs in A-DNA. (a) Stick rendering of the crystal structure of 3/6 ZP with N, blue; O, red; P, orange and C is green for Z and yellow for P; standard pairs are in orange. (b) End view of the same structure.

The properties of the 3/6ZP DNA structure containing multiple consecutive P:Z nucleobase pairs were analyzed using 3DNA. 

The first significant finding was that consecutive non-natural P:Z nucleobase pairs can be accommodated in canonical A- and B-helical forms of DNA. As in the B-form, the P:Z pairs exhibit on average a wider major groove than that of the G:C structure used for comparison with an average major groove width of 18.9 Å versus 18.0 Å on average. In both structures, P:Z formed three hydrogen bonding interactions with distances between heteroatoms typical of those in natural nucleobase pairs. The P:Z pairs in B-form DNA were sheared as were equivalent G:C pairs (Figure 4), whereas those in the A-form structure were not sheared, similar to G:C pairs in A-form DNA (Figure 7). Two different types of stacking interactions for PP/ZZ dinucleotide steps were identified in the 2P vs the 3/6 ZP structures. In the B-form 2P structure, consecutive Zs are stacked in much the same manner as natural nucleobases in a shifted arrangement, whereas in the A-form 3/6ZP structure, the nitro group of Z stacks above the ring of the adjacent Z. This stacking arrangement involves a sliding motion comparable to that seen for G:C steps in A-form DNA (Figure 8).

Stacking interactions between adjacent nucleobase pairs contribute to helix stability, and ample precedent exists for using QM calculations to obtain estimates of stacking energies for adjacent Watson–Crick nucleobase pairs. We and others have therefore used these methods to demonstrate that stacking P:Z nucleobase pairs is energetically preferred to stacking G:C nucleobase pairs by approximately 2.0 kcal/mol primarily because of favorable electrostatic interactions between the electron-deficient Z ring and the π-electrons of the adjacent P nucleobase (Figure 2). Perhaps more importantly for
duplex stability and conformational properties for DNA containing multiple consecutive P:Z nucleobase pairs, high-level QM calculations suggest that the PP/ZZ dinucleotide (i.e., two consecutive P:Z pairs) can adopt two different structural forms (Figure 8), which we have termed "slide" and "shift" conformers because one of the two nucleobase pairs is displaced along the axis corresponding to slide or shift, respectively.21 Comparison with X-ray crystal structures shows that the slide conformer is similar to what is observed for the stacking of PP/ZZ dinucleotides in A-form DNA, whereas the shift conformer resembles that observed in B-form DNA. The calculated energy difference between the two structures (1.5 kcal/mol) suggests that the "slide" conformer, which features "staggered" stacking of the nitro groups (Figure 8), is more stable, but both conformers are accessible at room temperature and above. Similar QM studies for the GG/CC dinucleotide suggest a greater energetic preference for the slide conformer, which is consistent with the experimental finding that duplexes containing only G:C nucleobase pairs prefer to exist in the A-form structure.30 Of course, access to viable cells containing an expanded genetic alphabet requires not only that DNA duplexes containing multiple P:Z nucleobase pairs adopt canonical helical structures but also that PZ-containing sequences can form specific interactions with proteins, such as transcription factors and repressors, in the major groove.40 As a consequence, redesigning DNA-binding proteins to recognize AEGIS nucleobases requires an understanding of the electrostatic properties of the P:Z nucleobase pair within the duplex.

■ MODELING THE STRUCTURE AND DYNAMICS OF AEGIS DNA

Elucidating the intrinsic properties of individual AEGIS nucleobases using QM calculations is a necessary element of developing simplified models, such as force fields, which relate changes in molecular geometry to potential energy. Given the intrinsic biological importance of understanding how DNA sequence might impact conformational preferences, flexibility, and the motional properties of the double helix, high-quality force field parameters now exist to model all of the standard Watson–Crick nucleobases in their lowest energy tautomeric form.42–44 In addition, there have been systematic studies on the interaction of DNA with counterions,45 such as Na+, and modeling the electrostatic properties of this highly charged biopolymer using classical potential energy functions.46 This very large body of work, together with long time scale molecular dynamics (MD) simulations, has resulted in a detailed understanding of duplex DNA structure in water and the dynamical motions that mediate conformational transitions.47

In contrast, there have been almost no reports of simulations of duplex DNA built from AEGIS nucleobases, perhaps because of a lack of well-tested force field parameters for these novel nucleobases. To understand the molecular basis for the unique structural features seen in the crystal structures of PZ-containing DNA duplexes, we therefore developed parameters for these two AEGIS nucleobases assuming that they exist only in their lowest energy tautomeric form (Figure 1). With these in hand, we were then able to examine the motions and conformational properties of the 3/6ZP oligonucleotide in water over a 50 μs time scale. This calculation showed that the presence of the six consecutive P:Z nucleobase pairs gave rise to a duplex that exhibited structural features associated with both A- and B-form DNA. For example, the 3/6ZP oligonucleotide featured a wider major and narrower minor groove (average values of 27 and 13 Å, respectively) over the course of the MD simulation than in the A-form helix observed in the X-ray crystal structure (values of 19 and 16.5 Å, respectively).30 The differences in groove width for the 3/6ZP oligonucleotide in water are, of course, one of the defining features of the B-form double helix. On the other hand, the range of values of structural measures associated with PP/ZZ dinucleotide steps are those expected for A-form duplex DNA.21 This conformational behavior is likely associated with adjacent P:Z nucleobase pairs interconverting between the "slide" and "shift" conformers identified by QM calculations. In addition, when compared with the dynamical behavior of a "control" DNA duplex in which P:Z nucleobase pairs in the 3/6ZP oligonucleotide are all replaced by G:C, the PZ-containing DNA duplex in water accesses a larger number of conformations over the course of the MD simulations (Figure 9).

These observations contrast sharply with conclusions from an MD simulation of a 15 bp DNA duplex containing only a single P:Z nucleobase pair,15 which suggested that the presence of the AEGIS nucleobases had little impact on the duplex structure. These MD simulations do, however, suggest that bacteria will be able to tolerate the inclusion of P:Z nucleobase pairs in their genome given that PZ-containing DNA duplexes can adopt both A- and B-form DNA. Thus, the presence of well-defined major and minor grooves will permit interactions with DNA-binding proteins involved in controlling transcriptional replication. In addition, the ability to adopt A-form structures will facilitate DNA replication in engineered DNA polymersases, and the similarity in hydrogen bond interaction energies for G:C and P:Z nucleobase pairs makes formation of transcription "bubbles" energetically feasible for PZ-containing DNA. Perhaps most importantly, these simulations show that the DNA duplex can be maintained in water even for sequences composed of multiple consecutive P:Z nucleobase pairs.

Figure 9. Representative PZ and GC structures from MD simulations. Stick models are shown for the PZ oligonucleotide in an extended conformation (far left), PZ in an A-like conformation (center), and GC in a B-like conformation. Views are shown parallel (upper panels) and perpendicular (lower panels) to the helical axes. In each rendering, N is blue, O red, P orange, and C light gray. Taken from ref 19 and used with permission.
CONCLUSIONS AND OUTLOOK

This Account has provided an overview of the properties of AEGIS nucleobases and AEGIS DNA highlighting our most recent crystallographic and computational studies of PZ-containing DNA and relevant natural control sequences. Collectively, our studies provide the first comprehensive analysis of the structural and biophysical properties of artificial DNA containing multiple P-Z pairs in a six nucleotide generic alphabet. P-Z pairs exhibit structural features similar to those of G-C pairs in the same context in both B- and A-DNA crystal structures, suggesting that they would be expected to behave similarly in biological reactions. Notable differences include the unusual electrostatic and stacking properties of the P-Z nucleobase pairs, which we suggest contribute substantially to the observed dynamical differences between the duplex, 5’-CTTATPPPZZZATAAG, and the GC control. Through optimization of DNA polymerase by directed evolution, it is possible to faithfully and efficiently replicate DNA including P:Z pairs in PCR reactions. Thus, all of the studies to date support the possibility that P-Z pairs expand the genetic alphabet and in so doing the fundamental properties of the DNA.

AUTHOR INFORMATION

Corresponding Author

*Tel: 1(317)278-8486, fax: 1(317)274-4686, e-mail: mgeorgia@iu.edu. ORCID

Nigel G. J. Richards: 0000-0002-0375-0881

Millie M. Georgiadis: 0000-0003-2976-4576

Notes

The authors declare no competing financial interest.

Biographies

Nigel G. J. Richards is Professor of Biological and Organic Chemistry at Cardiiff University, UK. He is also a Research Fellow at the Foundation for Applied Molecular Evolution based in Alachua, FL.

Millie M. Georgiadis is Associate Chair and Professor in the Department of Biochemistry and Molecular Biology at Indiana University School of Medicine and has a joint appointment in the Department of Chemistry and Chemical Biology at Indiana University–Purdue University Indianapolis.

ACKNOWLEDGMENTS

The authors would like to acknowledge the efforts of a number of talented postdoctoral and graduate student researchers who contributed to both the experimental and computational studies (especially Robert Molt and Isha Singh) outlined herein as well as Dr. Steven Benner (fFAME), who created AEGIS. This work was supported in part by an NIH grant R01 GM055026 to M.M.G. and funds provided by Cardiiff University and Indiana University–Purdue University Indianapolis to N.G.J.R.

REFERENCES


