In Silico Footprinting of Ligands Binding to the Minor Groove of DNA

Nahoum G. Anthony,[†,‡] Guillaume Huchet,[†] Blair F. Johnston,[†] John A. Parkinson,[‡]
Colin J. Suckling,[‡] Roger D. Waigh,[†] and Simon P. Mackay*[,†]

Department of Pharmaceutical Sciences, University of Strathclyde, 27 Taylor Street, Glasgow G4 0NR,
Scotland, and Department of Pure and Applied Chemistry, University of Strathclyde, 295 Cathedral Street,
Glasgow G1 1XL, Scotland

Received April 26, 2005

The sequence selectivity of small molecules binding to the minor groove of DNA can be predicted by "in silico footprinting". Any potential ligand can be docked in the minor groove and then moved along it using simple simulation techniques. By applying a simple scoring function to the trajectory after energy minimization, the preferred binding site can be identified. We show application to all known noncovalent binding modes, namely 1:1 ligand:DNA binding (including hairpin ligands) and 2:1 side-by-side binding, with various DNA base pair sequences and show excellent agreement with experimental results from X-ray crystallography, NMR, and gel-based footprinting.

INTRODUCTION

Following the publication of the human genome, there is a major research effort aimed at the design of DNA-binding compounds that target genes or their promoter regions.[1−9] Through their selective association with a given DNA sequence, such ligands may prevent protein binding and treat diseases that result from aberrant gene expression. The genetic code is expressed through the major and minor grooves of DNA by reading the hydrogen bonding capacity of the base pair edges on the groove floors, and a number of groups have been designing selective ligands called lexitropsins or polyamides on this basis, particularly to bind in the minor groove. There are a number of excellent reviews[10−13] covering the current status of research in this general area.

The development of lexitropsins arose from the observation that two natural antibiotics, netropsin and distamycin (Figure 1), bound to A and T containing regions of the minor groove by a combination of hydrogen bonding with the bases on the groove floor and a natural isohelicity with the groove itself. G and C containing regions were excluded because of the steric repulsion between the C−H of the *N*-methylpyrrole (Py) on the inner face of the molecule and the exocyclic G-NH_2, which protruded from the minor groove floor.[14,15] To overcome this restriction, Py was replaced with *N*-methylimidazole (Im), where the heterocyclic N could accommodate G-NH_2 by hydrogen bonding.[16,17] Unfortunately, while an Im moiety could bind to G:C base pairs, it could also accommodate A:T regions. The significant breakthrough in the field came with the observation that a number of lexitropsins could bind in the minor groove as a 2:1 complex, in a side-by-side fashion with the heterocyclic rings stacking against each other.[18] As a result, a number of lexitropsins were prepared that could discriminate not only G:C from A:T but also G:C from C:G and A:T from T:A base pairs,[19,20] although the latter discrimination resulted in a considerable loss of affinity for the site through the use of a hydroxypyrrole heterocycle.[7]

More recently, it has become apparent that while hydrogen bonding to the groove floor provides specificity for particular sequences, the major driving forces for association between lexitropsins and DNA are lipophilic and hydrophobic, in particular interactions with the sugar moieties that comprise the groove walls.[21,22] In this context, we have been preparing a large number of new lexitropsins made up from new heterocyclic groups that seek to recognize both the hydrogen bonding capacity of the groove floor to achieve specificity and to exploit the lipophilic nature of the groove walls to enhance affinity.[2−4,23,24] Indeed, with the new lexitropsin thiazotropsin A (Figure 1), we have shown that lipophilic substituents can be instrumental in determining sequence specificity; in addition to influencing affinity, such groups can extend reading frames through their ability to modify the nature of side-by-side association.[24]

When developing effective DNA minor groove binding lexitropsins, the practical rules devised by Dervan's group[20] are useful guidelines. To also predict the sequence selectivity and affinity of any new ligand by including van der Waals and other forces would be invaluable for prescreening the large number of potential compounds arising from a synthetic program. Inevitably, any modern drug-design program is limited not only by the synthesis of new ligands (for which parallel automated combinatorial methods are perfectly tailored) but also in the ability to assess the effectiveness of the drug-design paradigms. Over the past 25 years, much progress has been made in the field of docking-based high throughput screening with protein targets, yet only one modeling study has explored a virtual screening method with sequenced DNA, using a single ligand (berenil) that required skilled computer programming and operator intensive methodologies.[25] We have developed an extremely efficient and simple automated docking procedure that we have called in

* Corresponding author phone: 00 44 141 5482866; fax: 00 44 141 5526443; e-mail: simon.mackay@strath.ac.uk.
[†] Department of Pharmaceutical Sciences.
[‡] Department of Pure and Applied Chemistry.

10.1021/ci050153b CCC: $30.25
Published on Web 10/05/2005

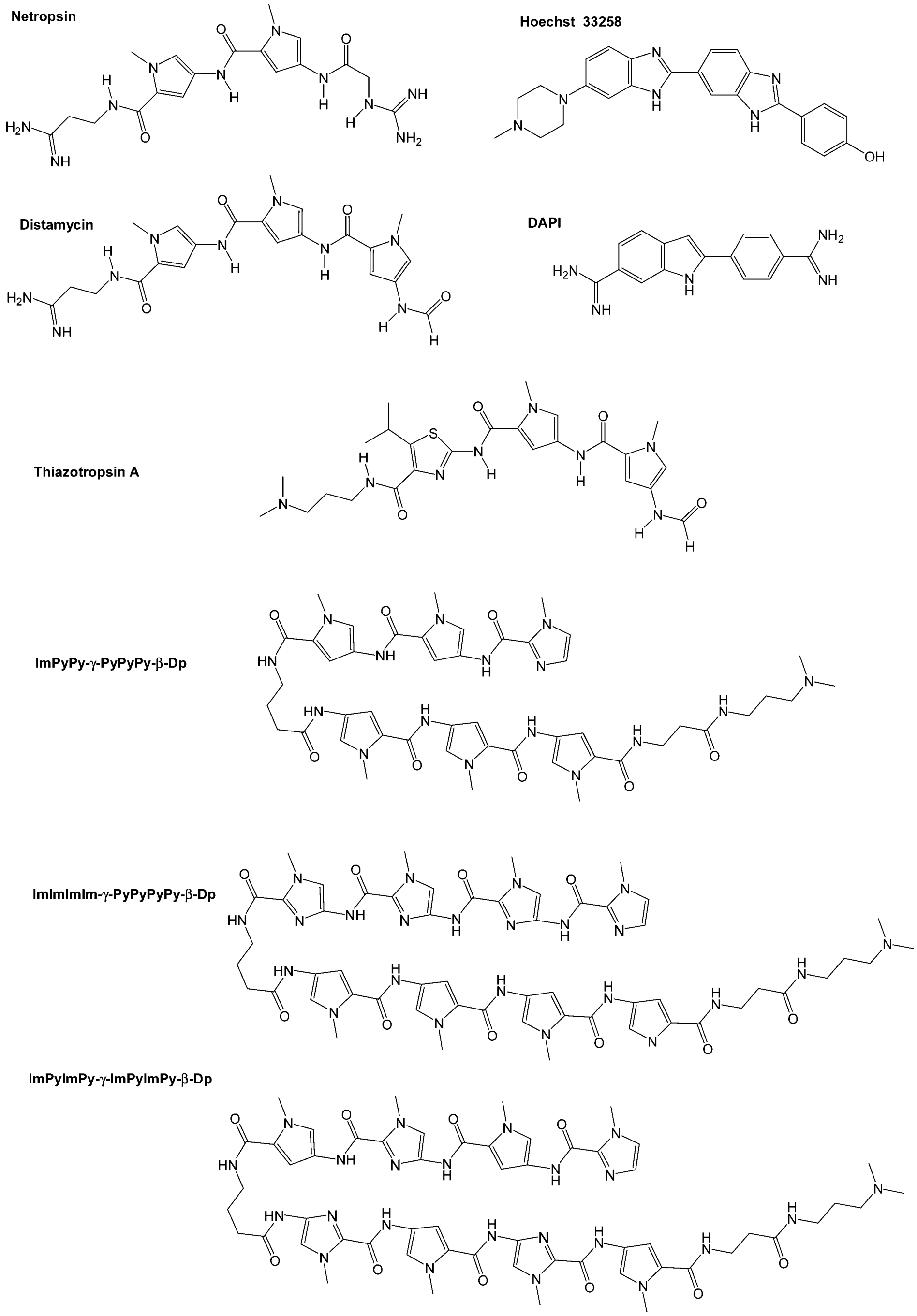

Figure 1. Structures of common lexitropsins that bind to the minor groove of DNA.

silico footprinting (ISF), which can generate thousands of structures in a single trajectory for evaluation. We have shown that a simple scoring function applied to the structures in the minimized trajectory can give a good indication of sequence selectivity for the ligand concerned. This was crucial if our method was to offer a rapid prescreening process for compounds proposed in our synthetic program.[2,3,23] Here we report method development and application of ISF to 1:1 and 2:1 side-by-side ligand:DNA binding.

RESULTS AND DISCUSSION

Our aim was to automate the movement of a ligand along any DNA sequence and to identify the preferred binding site for that ligand. It was important that the ligand should not move as a rigid body through a translational movement alone, but be allowed to adapt to the topography of the minor groove, which itself changes according to the sequence. Docking the ligand at the start of the sequence and applying a pulling force in the form of a quadratic distance constraint between the leading end of the ligand and the end of the DNA using molecular dynamics achieved this (Figure 2, for a simulation movie with netropsin, see Supporting Information). To prevent the trailing end of the ligand straying out of the groove during the simulation, a low antistraying distance restraint was required between the trailing end and the starting base pair. Providing that the force is substantially smaller in magnitude than the leading end pulling force, a net force in the desired direction is achieved. An evaluation of the sequence affinity was obtained by minimization of the coordinates stepwise along the trajectory once the

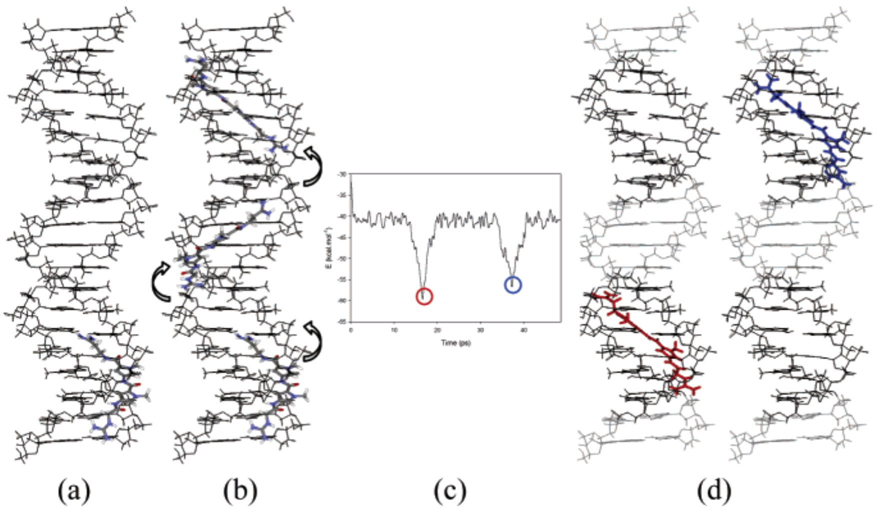

Figure 2. (a) The ligand is docked at the start of the sequence (b) a pulling force is applied to automate movement along the groove (c) structures within the trajectory file are minimized and potential energy plotted against time (d) the lowest energy structures in the profile are examined for sequence preference. For movie simulations, see the Supporting Information.

simulation was complete, followed by computation of the potential energy of the ligand complex at each step against distance travelled; lower potential energy regions indicated where the ligand preferred to bind. Essentially, the scoring function for the thousands of structures generated is energy based and relies on the nonbonded interactions between the ligand and the minor groove of the DNA to identify the preferred binding site. Given that numerous authors have shown that binding is governed considerably by hydrophobic and lipophilic interactions between the ligand and the DNA,[21,22,26,27] and with our emphasis being on the development of a fast and efficient automated screening technique, we selected this means of scoring poses based on electrostatic and van der Waals contributions and the internal energy of the ligand. Such atom-based fitness functions are used in GOLD,[28] while Standard Precision Glide measures the interaction energy of each ligand with Coulomb and van der Waals grids.[29] We found that the only tuning necessary to give reliable and efficient scoring was to generate electrostatic potential charges on the ligand using the AM1 method, rather than using empirical force field charges. A 6−9 Leonard Jones potential was used, the energy cutoff was 14 Å, and standard Coulombic interaction parameters defined the electrostatic contribution with a distance dependent dielectric of $4r_{ij}$. Ligand internal energies employed default force field input for bonded and nonbonded energies, and the scoring output was given by the equation

$$E_{\text{score}} = E_{\text{ligand}} + E_{\text{nonbonded}}$$

We accept that this nonrigorous method of evaluation does not take into account more stringent thermodynamic parameters—our emphasis was on a virtual screening scoring function for the thousands of structures generated. Such detailed analyses on selected structures can be performed at a later date, in the same way that any protein-based screening results are further refined once hits have been identified.

To develop a technique with extended tracts of DNA, it was clear from the start that water could not be included explicitly in the simulations. Employing periodic boundary conditions on large sequences of DNA over extended simulation times, followed by the minimization of each frame in the trajectory was too computationally expensive. Simulations were therefore performed *in vacuo* using a distance-dependent dielectric of $1r_{ij}$. When subjecting DNA to dynamics simulations *in vacuo*, problems are often encountered with charges on the sugar−phosphate backbone: repulsive nonbonded interactions between neighboring phosphate groups often result in untenable structures through strand uncoupling and unravelling.[30,31] To overcome these shortcomings, charges are often reduced in magnitude or removed entirely and a variety of restraints or constraints applied.[30,32−36] We found that pulling a ligand along the minor groove, even after applying constraints or restraints to the DNA, resulted in ligand trapping through electrostatic interactions at various positions along the groove and thereby prevented a full exploration of the DNA sequence by the ligand. Increasing the pulling force to overcome this electrostatic trapping resulted in the ligand leapfrogging along the groove and not sampling the entire groove sequence. To overcome electrostatic trapping, simulations were performed in the absence of all charges on fixed DNA.[32] Charges were reassigned when structures in the trajectory files were minimized to obtain a representative potential energy plot that included electrostatic contributions.

1:1. Ligand:DNA Complexes (Hairpin Excluded). The technique was initially developed using experimentally determined structures of ligand/DNA dodecamer complexes. Fixing the DNA to its experimentally determined coordinates and applying pulling and antistraying force constants (see Methods) produced a trajectory for the length of the groove for netropsin, distamycin (Figure 3a,b), and Hoechst 33258 (data not shown). Analysis of the potential energy against time showed that the optimum energy score (or ‘footprint’) was in excellent agreement with the original experimentally determined ligand−DNA complexes, with RMS values < 0.5. (RMS values ≤ 0.5 with the experimentally observed binding mode are considered high in ranking accuracy when using and evaluating docking algorithms.)[37]

We observed that allowing DNA some flexibility by tethering rather than constraining gave poor sequence affinity predictions, as the conformational energy of the DNA gave the major contribution to the overall energy of the complex,

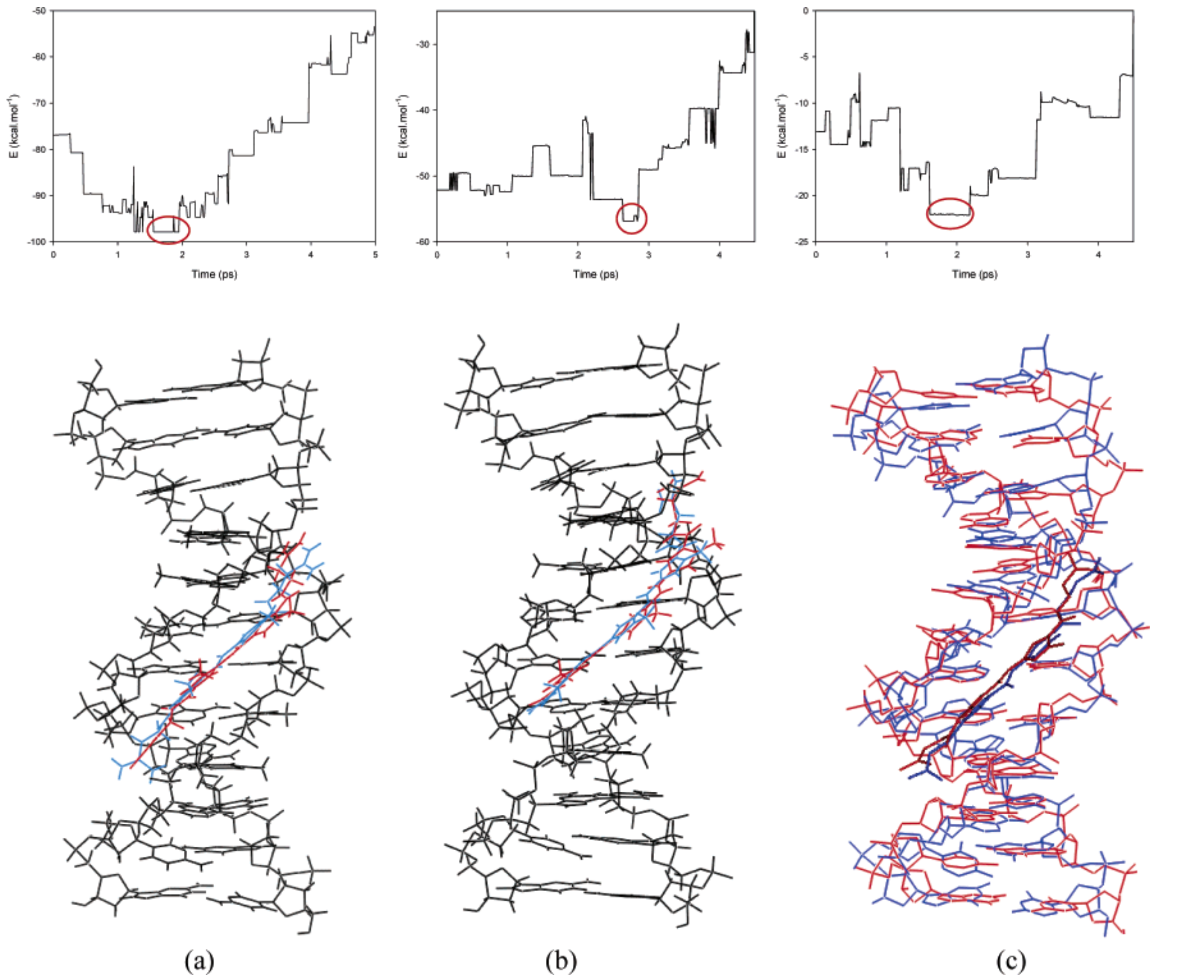

Figure 3. (a) ISF of netropsin using experimentally determined coordinates (top) and superimposition of the original crystal (blue) and structure found with ISF (red) RMS = 0.43. (b) ISF of distamycin using experimentally determined coordinates (top) and superimposition of the original crystal (blue) and structure found with ISF (red) RMS = 0.42. (c) ISF of netropsin using standardized B-DNA coordinates (top) and superimposition of the original crystal (blue) and structure found with ISF (red).

and swamped the contribution of the ligand and its interaction with the groove. Using tethered DNA, the energy contribution by the ligand−DNA interaction is lost within the potential energy contribution from the DNA because it is no longer fixed and makes the identification of a footprint impossible. Thus, comparison between the energies of the conformations of a complex along its trajectory was more appropriate when the DNA was fixed and the change in energy was due only to the position of the ligand, producing a relative selectivity map of the ligand over the given sequence.

Achieving successful sequence footprinting for a ligand exploring the minor groove of a section of DNA that uses the coordinates from the experimentally determined structure has its limitations. There are a restricted number of DNA sequences for which experimental coordinates are available, yet the versatility of the method relies upon its applicability to footprinting any DNA sequence. Significantly, we were able to reproduce the results for netropsin, distamycin, and Hoechst 33258 with excellent correlation when using the same sequences built using standard B-DNA structural parameters (Figure 3c)[38,39] (minimized for 100 iterations prior to footprinting to remove bad contacts arising from the original standard parameter conformation), which demonstrates that we do not have to rely on DNA structures determined by crystallography or NMR.

To establish whether our method was sensitive to subtle changes in sequence, and whether it had any DNA length limitations, we applied it to netropsin and distamycin with standardized DNA that contained the two binding sites, 5′-AAATTT and 5′-TATATA, within random sequences of Gs and Cs in 80 and 100 base pair tracts (See simulation movie for netropsin in the Supporting Information). Netropsin and distamycin bind to AT regions of DNA in a rank order, with preference for a 5′-AT step over a 5′-TA step.[40] ISF on these extended sequences showed excellent agreement with the experimental data for both polyamides (Figure 4). Simulations with more than 80 base pairs did not reach completion because of software memory limitations; longer sequences of DNA have to be analyzed by splitting the trajectory into multiples of 80 or less base pair sequences. We applied the same methodology to the nonpolyamide lexitropsins Hoechst 33258 and DAPI (Figure 1) with extended DNA sequences and again demonstrated that single or preferential binding sites among related sequences could be reliably identified (Figure 4). Particularly impressive were the results for Hoechst 33258 with a DNA tract that contained five potential binding sites with disparate preferences for the ligand as revealed by gel footprinting; ISF reproduced these differential affinities in the correct order (Figure 5).

Side-by-Side Ligand:DNA Complexes. When lexitropsins bind to DNA in a side-by-side fashion, the minor groove

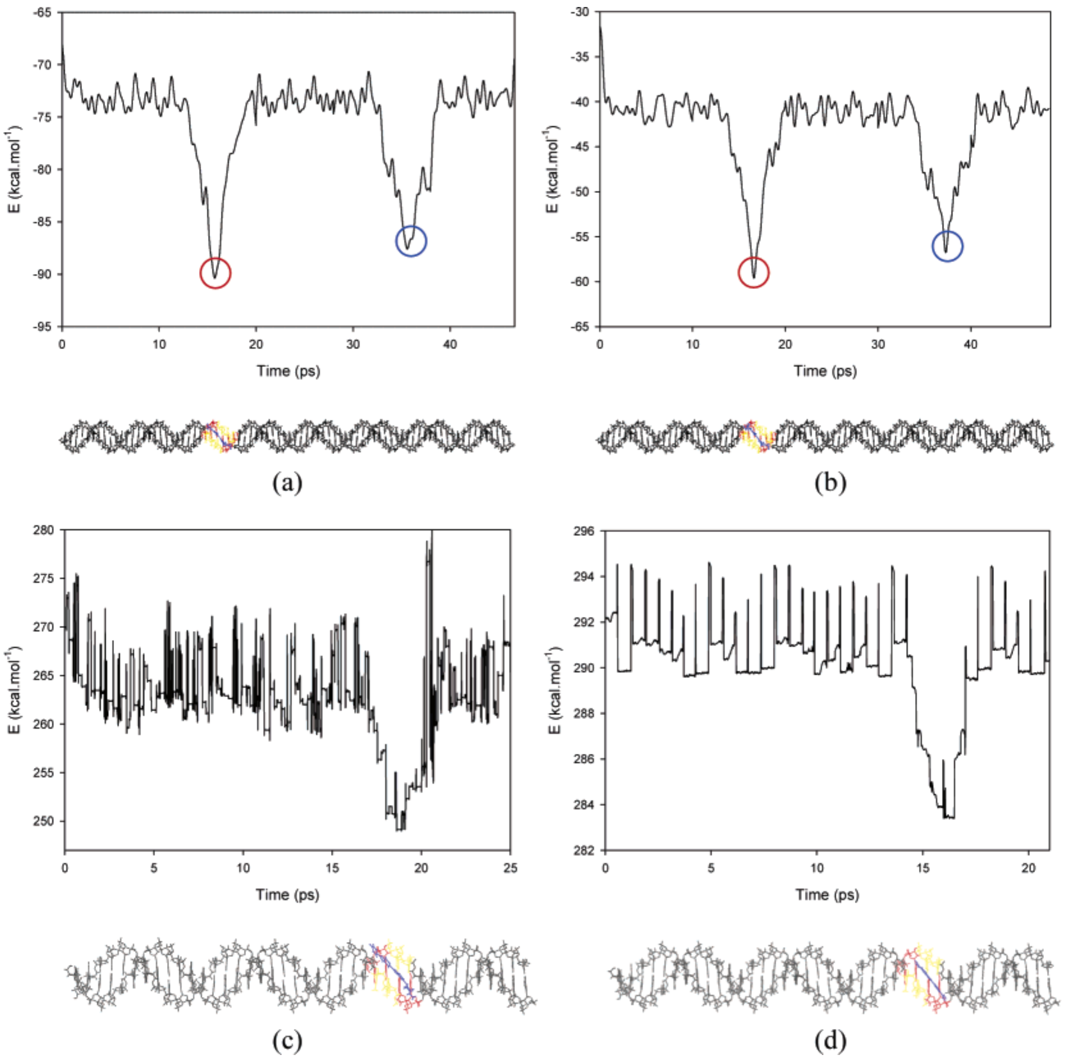

Figure 4. (a) ISF of netropsin on an 80 base pair sequence containing the two target sites 5′-AAATTT and 5′-TATATA. The minima are circled in red (5′-AT step) and in blue (5′-TA step). (b) ISF of 1:1 distamycin on an 80 base pair sequence containing the two target sites 5′-AAATTT and 5′-TATATA. The minima are circled in red (5′-AT step) and in blue (5′-TA step). (c) ISF of 1:1 Hoechst 33258 on a 40 base pair sequence containing the target site 5′-AATT. (d) ISF of 1:1 DAPI on a 40 base pair sequence containing the target site 5′-AATT.

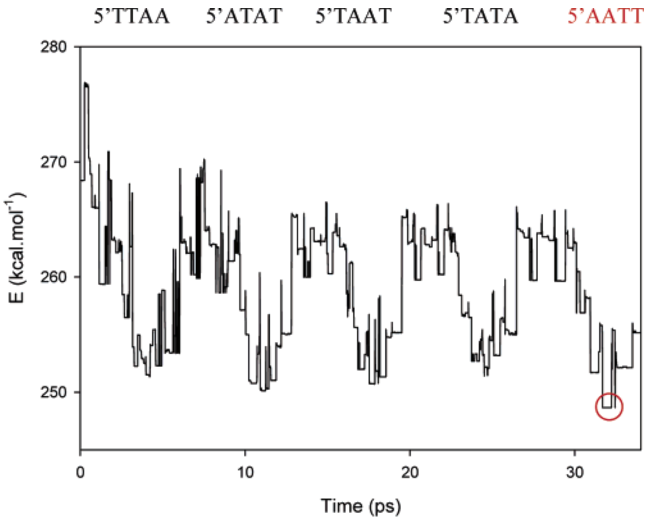

Figure 5. Hoechst with five sites with comparison of the order of binding from[40] which is AATT>TAAT=ATAT>TTAA=TATA.

widens significantly in order to accommodate the two ligands. Proximal to the binding site, the minor groove re-establishes the narrower dimensions characteristic of B-DNA. From an ISF perspective, the force required to move side-by-side ligands along the minor groove would also need to effectively widen the groove as the ligands progressed while at the same time maintaining the adjacency of the ligands. When restraints are removed from the DNA to accommodate the widening in groove dimensions as the ligand progresses, helical unwinding and strand separation takes place.[30,32−36] Having established that for 1:1 ISF fixing the DNA coordinates gave optimum sequence affinity predictions, to establish a similar protocol for side-by-side binding required the assembly of DNA sequences with wide minor groove dimensions (wide groove or WG-DNA, see Methods) through which the ligands could be pulled. We determined the general parameters for such DNA so that “wide-groove” (WG) DNA of any length and sequence could be constructed. Parameters for both wide groove GC and AT tracts were calculated from experimentally determined structures (pdb reference 334D[41] for GC and pdb reference 378D[42] for AT) (see Methods and Figure 6). To ascertain whether ISF could sequence 80 base pair tracts successfully for 2:1 binding, we subjected distamycin to ISF in a side-by-side binding mode employing a WG-80 base pair sequence that contained mixed Gs and Cs and a 5′-AAATT-3′ target in its core (see movie in the Supporting Information). The distamycin dimer was docked

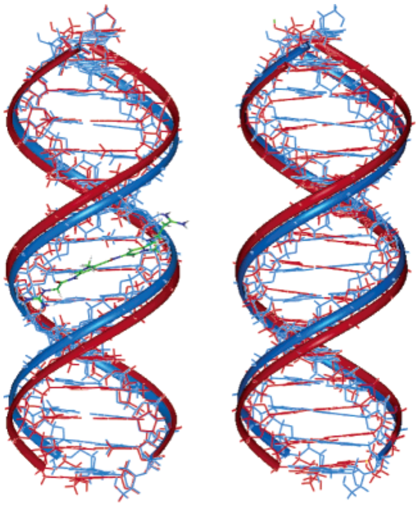

Figure 6. Effect of building DNA using wide-groove parameters on the minor groove width of a poly AT sequence (left) and a poly GC sequence (right). Parametrized wide groove DNA is colored in red and superimposed to canonical DNA of the same sequence (blue). The ribbons show the location of the phosphate backbone of DNA. A ligand has been docked on the poly AT sequence to help locate the minor groove.

at the starting base pair using the conformation taken from crystallographically determined structures. Through a combination of strong and weak constraints to pull the dimer along the groove while maintaining its relative adjacency (see Methods), we showed that the experimentally determined 5′-AAATT-3′ binding site[18] was uniquely identified within the 80 base pair sequence (Figure 7).

It is not altogether too surprising that A/T binding sites can be identified within extended GC tracts, given the steric repulsion that the exocyclic G-NH_2 groups offer on the minor-groove floor. However, the method's ability to discriminate between different A/T sequences using a simple potential energy function is impressive (Figures 3 and 4). The real challenge for ISF was to sequence more complex binding sites having all four base pairs with a number of side-by-side and hairpin ligands that incorporated G:C, I:C, and A:T-recognizing motifs. Dervan has shown that hairpin structures overcome the probability of two separate ligands binding simultaneously side-by-side to the target sequence.[43] Covalently linking the *C*-terminus of one ligand with the *N*-terminus of the adjacent ligand, usually with a γ-aminobutyric acid linker (γ) (Figure 1), forces aromatic stacking between the heterocyclic units in a predefined manner and enables G/C and A/T sites within a binding sequence to be selectively targeted. We initially performed ISF with such a representative ligand, which was able to accommodate a GC step within its target sequence. Winston and co-workers demonstrated that the hairpin ligand ImPyPy-γ-PyPyPy-β-Dp (Figure 1) selectively bound to a 5′-TAACAA-3′ site in preference to other sites such as the 5′-TAACA-3′ and 5′-AACAA-3′ sites.[44] We constructed a 48 base pair WG-DNA sequence incorporating these three binding sites separated by 5′-GGCCGGCC-3′ spacers, and ISF successfully identified the binding sites in the correct order of affinity (Figure 8).

Dealing with multiple G:C steps within binding sites was equally successful; the hairpin ligands ImImImIm-γ-PyPy-PyPy-β-Dp and ImPyImPy-γ-ImPyImPy-β-Dp (Figure 1) target the 5′-TGGGGA and 5′-TGCGCA sites, respectively.[45] When the latter site was incorporated within an 80 base pair WG-DNA sequence of mixed Gs and Cs, it was successfully identified (Figure 9a). More impressively when both sites were included in an 80 base pair WG-DNA sequence separated by 5′-GGGCCC spacers, along with the extra site 5′-TGGCCA that is subtly different from the target regions, the binding sites were successfully discriminated by both ligands (Figure 9b,c).

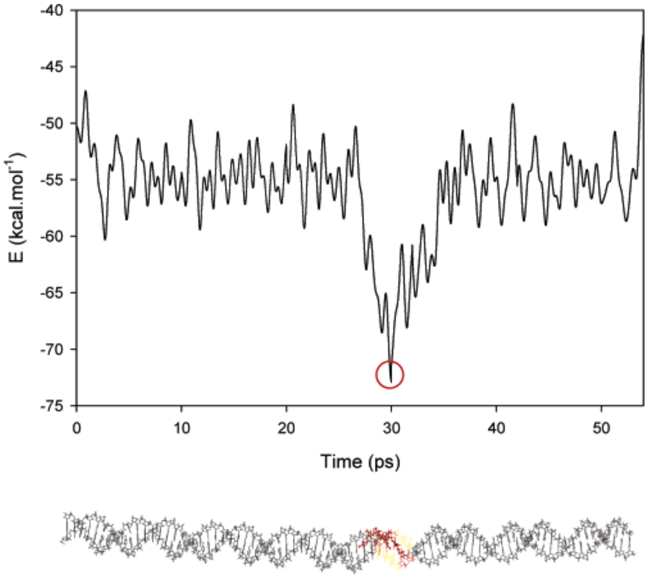

Figure 7. ISF of 2:1 distamycin on an 80 base pair sequence containing the target site 5′-AAATT. The minimum is circled in red.

With the recognition that our method could identify binding sites that incorporated all four natural base pairs with a variety of polyamides and lexitropsins that bound in 1:1, 2:1, or as hairpin complexes, we finally tested its ability to distinguish between G:C and I:C base pairs. The latter motif, while having the cytosine base in common, only has two Watson−Crick hydrogen bonds; inosine (I) is missing the exocyclic $-NH_2$. We have recently reported that thiazotropsin A preferentially binds to the sequence 5′-ACTAGT-3′,[46] and NMR spectroscopy has provided the details of its binding in a 2:1 complex within the sequence 5′-CGACTAGTCG-3′.[24] Interestingly, the bulky isopropylthiazole heterocylic unit within thiazotropsin A ensures a longer six base pair reading frame than is normally associated with such tricyclic lexitropsins is targeted. Indeed, the abnormal phase-shift between the stacked aromatic units that is observed in its binding mode means that Dervan's rules are not strictly obeyed for this lexitropsin. Impressively, ISF on thiazotropsin A correctly identified its 5′-ACTAGT-3′ binding site embedded within a mixed GC sequence, in its idiosyncratic 2:1 complex (Figure 10a). More recently, we have shown through NMR studies that binding between thiazotropsin A and the sequence 5′-CGCCTAGGCG-3′ is poorly resolved, while with the 5′-CGCCTAGICG-3′ site, the complex formed is similar in fashion to the 5′-CGACTAGTCG-3′ complex (data not shown). We attributed preferential binding at the site where I had replaced a G to the absence of an exocyclic $-NH_2$ that hinders binding of the dimethylaminopropyl tail in the minor groove. Particularly rewarding for us was that ISF of thiazotropsin A with an extended sequence containing

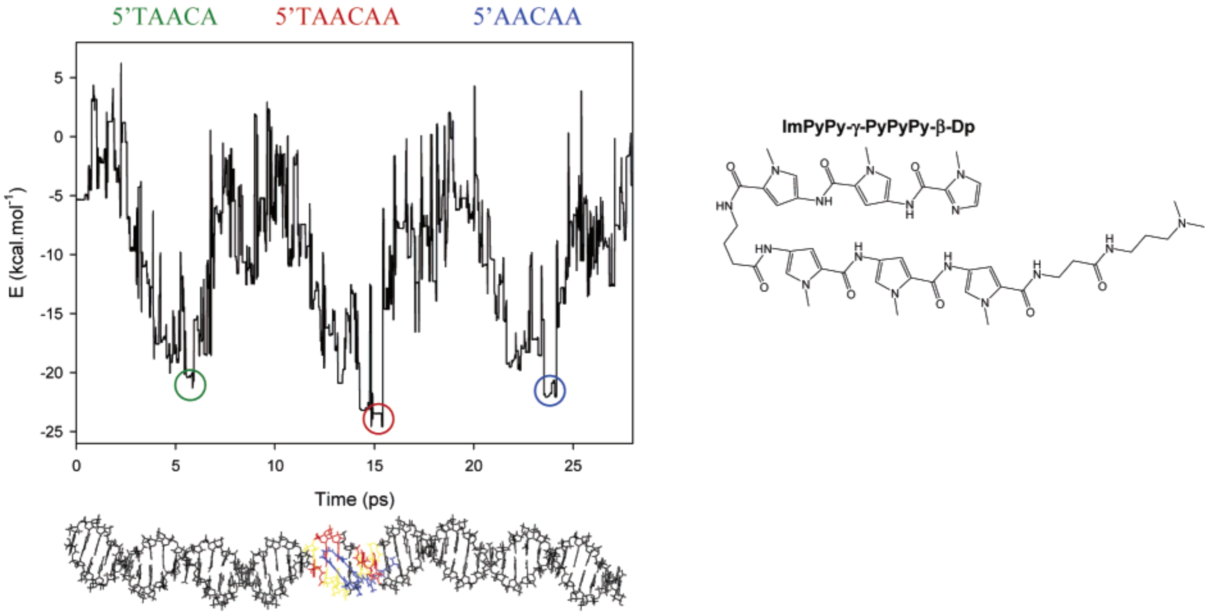

Figure 8. ISF of PyPyPy-γ-PyPyIm-β-Dp on a 36 base pair sequence of WG-DNA. The three binding sites are circled in red (5′TAACAA), blue (5′AACAA), and green (5′TAACA).

both 5′-CCTAGG-3′ and 5′-CCTAGI-3′ sites clearly identified the latter as its preferred binding site (Figure 10b).

Force-Field Dependency of ISF. To establish whether the identification of the experimental binding site using our virtual screening method was force-field independent we compared the systems using cff91, cff, cvff, and CHARMm force fields. Figure 11 shows representative comparisons for (a) netropsin on an 80 base pair sequence with two binding sites (5′-AAATTT-3′ and 5′-TATATA-3′), (b) thiazotropsin A with its preferred binding site 5′-ACTAGT-3′ embedded in a 40 base pair sequence, and (c) the ligand ImPyImPy-γ-ImPyImPy-β-Dp on a 40 base pair sequence containing the target site 5′-GCGC as well as the mismatch sequences 5′-GGGG and 5′-GGCC. We can clearly see the most robust force field that consistently identifies the correct binding site is cff91. In fact all the force fields used are able to distinguish between AT and TA binding sites for the ligands in question (Figure 11a). It appears that there is some loss of discernible scoring when Gs are admitted into the system. cff91, CHARMm, and cff all successfully identified the correct binding site for a side-by-side thiazotropsin A complex (Figure 11b). When multiple GC motifs are included to make three similar binding sites for the ligand ImPyImPy-γ-ImPyImPy-β-Dp to explore, only cff91 ranks the order correctly although all the force fields identify the binding sites themselves within the extended sequence (Figure 11c). For consistency, we therefore suggest the use of the cff91 force field for ISF.

In conclusion, we have developed an automated docking technique that allows us to scan extended DNA sequences and identify binding sites for small molecules. Over 5000 structures can be generated in less than an hour that fully covers any 80 base-pair sequence. Relative affinity for sequences can be determined by applying a simple scoring function to the minimized trajectories, which we have demonstrated by analyzing a wide selection of minor groove binding modes and sequences using different force fields. We have shown that our method can discriminate between subtle changes in the minor groove sequence incorporating all four bases and can be used to prescreen proposed ligands. The simplicity of the technique means that it will be accessible to the multidisciplinary workforce involved in genomic drug design and evaluation.

METHOD

InsightII, CHARMm, and CDiscover software (Accelrys Inc., San Diego, CA) was used to perform all calculations and molecule handling employing either the cff91,[47] cvff, cff, or CHARMm force fields. All simulations were performed using a 600 MHz, R14000, dual processor Silicon Graphics Octane2 or a dual processor Hewlett-Packard 3.2 GHz xw8200 workstation.

(i) Preparing the Simulation. The DNA coordinates for experimentally determined ligand–DNA structures were used without modification other than adding hydrogens and assigning force field potentials. ‘Standardized DNA’ refers to those sequences constructed using the InsightII software employing idealized B-DNA parameters. “Wide groove DNA” (WG-DNA) structures refer to those sequences constructed using our in-house developed parameters to accommodate side-by-side binding of ligands (Table 1). The charges used were the default force field charges for DNA and electrostatic potential (ESP) charges for the ligands, calculated using MOPAC[48] with the AM1 Hamiltonian. Prior to use, all standardized DNA and WG-DNA structures were minimized *in vacuo* using 100 steps of the conjugate gradient (CG) algorithm and a distant-dependent dielectric constant of $4r_{ij}$,[49] with the terminal base pairs restrained using a 10 kcal·mol^{-1} force constant.

The ligands were manually docked at the starting base pair of the minor groove and minimized with fixed DNA coordinates. The conformations of the ligands were based upon their coordinates taken from experimentally determined structures. Side-by-side binding included two possible scenarios: two individual lexitropsins associating side-by-side or hairpin lexitropsins whose molecular structure promotes side-by-side binding. For hairpin ligands, the hairpin loop was docked at the starting base pair.

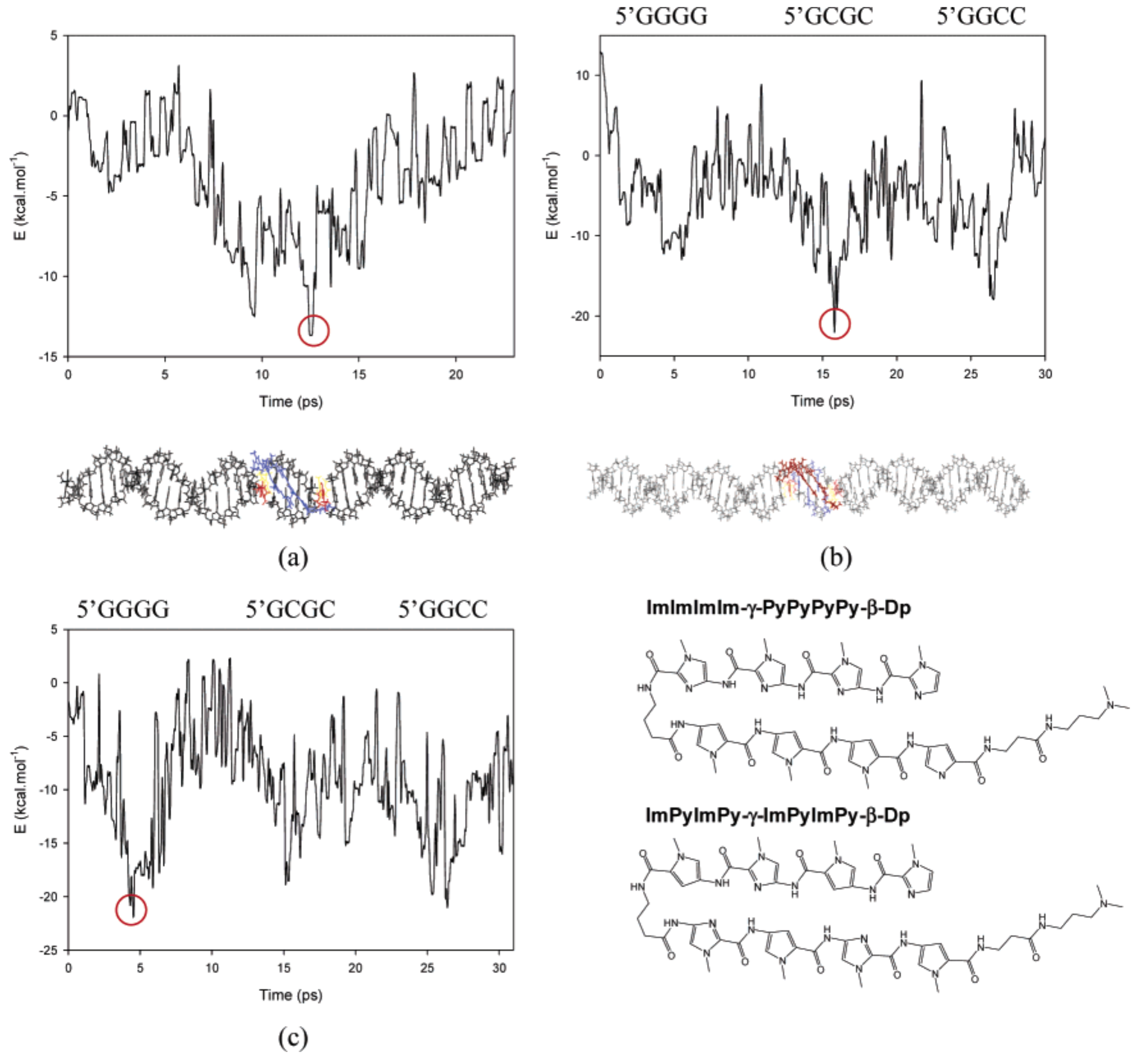

Figure 9. (a) ISF of the hairpin ligand ImPyImPy-γ-ImPyImPy-β-Dp on a 40 base pair sequence containing the target site 5′-GCGC. The minimum is circled in red. (b) ISF of the hairpin ligand ImPyImPy-γ-ImPyImPy-β-Dp on a 40 base pair sequence containing the binding sites 5′-GGGG, 5′-GCGC, and 5′-GGCC. The target site (5′-GCGC) correctly identified is circled in red. (c) ISF of the hairpin ligand ImImImIm-γ-PyPyPyPy-β-Dp on a 40 base pair sequence containing the binding sites 5′-GGGG, 5′-GCGC, and 5′-GGCC. The target site (5′-GGGG) correctly identified is circled in red.

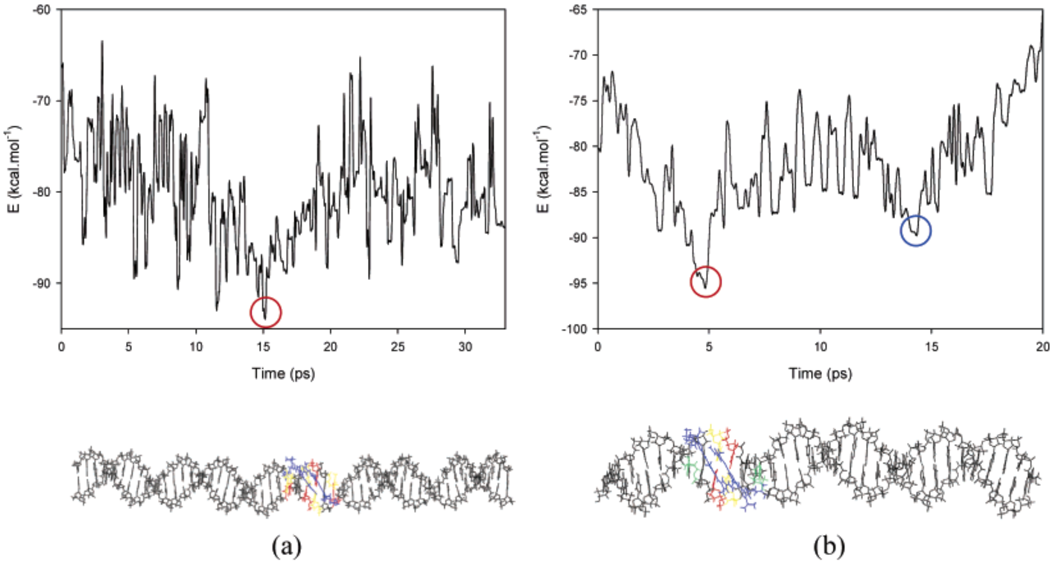

Figure 10. (a) ISF of thiazotropsin A on a 40 base pair WG sequence containing the target site 5′-ACTAGT (circled in red). (b) ISF of thiazotropsin A on a 40 base pair WG sequence containing the two binding sites 5′-CCTAGI (red) and 5′-CCTAGG (blue).

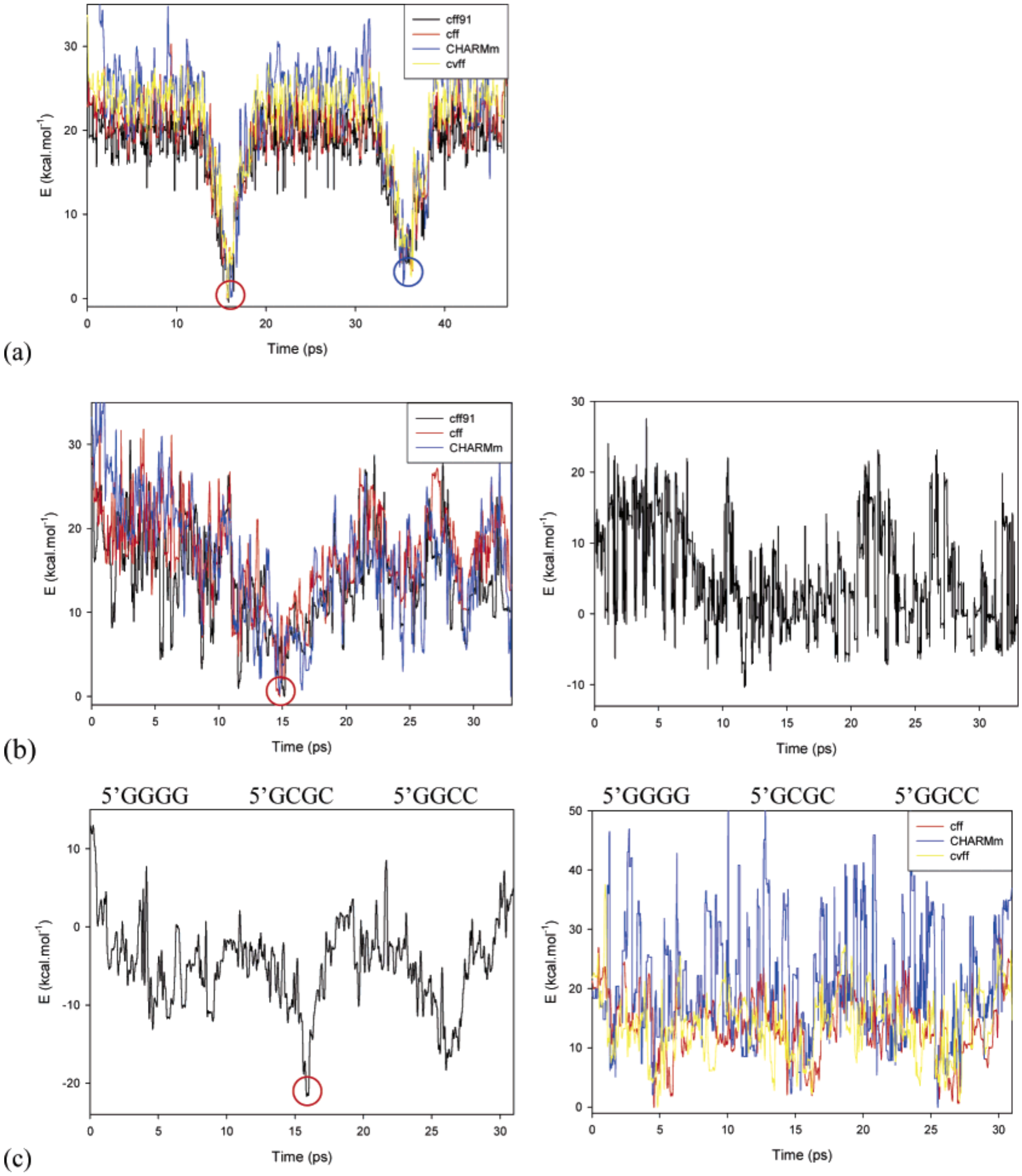

Figure 11. (a) ISF of netropsin on an 80 base pair sequence containing the two target sites 5′-AAATTT and 5′-TATATA. The minima are circled in red (5′-AT step) and in blue (5′-TA step). The four force fields tested identified the binding sites in their relative binding order. (b) ISF of thiazotropsin A on a 40 base pair sequence containing the target site 5′-ACTAGT (circled in red). Three of the force fields tested located the binding site (left), while the cvff force field did not (right). (c) ISF of the hairpin ligand ImPyImPy-γ-ImPyImPy-β-Dp on a 40 base pair sequence containing the binding sites 5′-GGGG, 5′-GCGC, and 5′-GGCC. The target site (5′-GCGC), correctly identified when using the cff91 force field, is circled in red (left), while the other three force fields used locate the three binding site but fail to rank them in the correct order.

Table 1. Parameters Used for Building DNA for ISF of 2:1 Ligand to DNA and Hairpin Complexes

step	X (Å)	Y (Å)	rise (Å)	tilt	roll	twist
AT or TA	0.21	−0.88	3.415	−0.9	2.7	−35.00
GC or CG	−0.12	−0.13	3.385	−2.2	6.9	−33.55

(ii) Running the Simulation. Simulations were performed *in vacuo* at 300 K employing a distant-dependent dielectric constant of $1r_{ij}$ with fixed DNA coordinates. Different sets of restraints were required depending on the geometry and stoichiometry of the system and are described below. All charges were removed from the DNA prior to dynamics simulations.

1:1 Ligand:DNA Complexes (Hairpins Excluded). Distance restraints were applied between the head of the ligand and the end of the DNA (referred to as the 'terminal' base pair) in order to create a pulling force on the ligand (Figure 12a). The dynamics simulation was run for the time required to simulate ligand movement from the starting to the terminal base pair with a 1 fs time step. To overcome tail straying, a low restraint ('antistraying') was required between the ligand tail and the starting base pair throughout the ISF simulation (Figure 11a).

2:1 Ligand:DNA Side-by-Side Complex. Distance restraints were applied between the head of the ligand closest to the terminal base pair (referred to as the leading end) and the terminal base pair in order to create a pulling force. Strong distance restraints at a fixed distance of 7 Å[42] were applied between the leading end and the neighboring leading

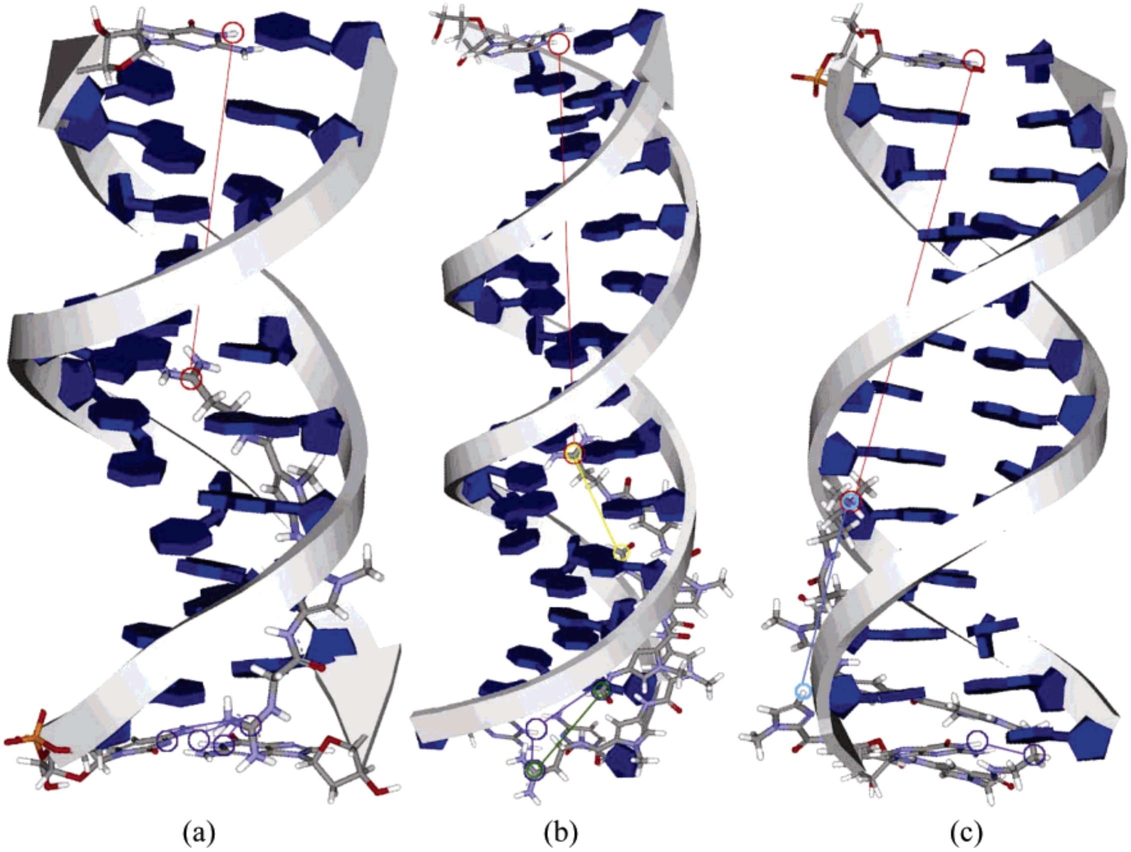

Figure 12. Examples of restraints applied to generate ISF simulations. (a) *Netropsin 1:1 binding*: a leading end pulling restraint (red) has an interatomic quadratic force constant of 2 kcal•mol^{-1}. Trailing end antistraying restraints (purple) have an interatomic quadratic force constant of 0.01 kcal•mol^{-1}. (b) *Distamycin 2:1 binding:* leading end pulling restraint (red) has an interatomic quadratic force constant of 5 kcal•mol^{-1}. Trailing end antistraying restraints (purple) have an interatomic quadratic force constant of 0.03 kcal•mol^{-1}. Adjacency restraints between leading ends (yellow) were quadratic and 500 kcal•mol^{-1} and between trailing ends (green) were flat-bottomed and 3 kcal•mol^{-1}. (c) *PyPyPy-γ-*PyPyIm-*β-Dp:* leading end pulling restraint (red) has an interatomic quadratic force constant of 5 kcal•mol^{-1}. Trailing end antistraying restraints (purple) have an interatomic quadratic force constant of 0.03 kcal•mol^{-1}. Adjacency restraints between leading ends (blue) were quadratic and 500 kcal•mol^{-1}.

end of the associated ligand to maintain adjacency. A low antistraying restraint was applied between both trailing ends of the ligands and the starting base pair of DNA. Additionally, a restraint was applied between the trailing ends themselves to reinforce cohesion of the system (Figure 12b).

1:1 Ligand:DNA Side-by-Side Hairpin Binding. Distance restraints were applied between the leading end of the ligand and the terminal base pair to ensure movement along the groove. Strong quadratic restraints were applied between both ends of the ligand to maintain a fixed distance in line with experimental data.[42] Finally, a low antistraying restraint was applied between the hairpin and the starting base pair (Figure 12c).

(iii) Analyzing the Simulation. All trajectories were obtained by saving the structure of the complex every 10 fs. Each structure in the trajectory file was then minimized *in vacuo*, using the same conditions employed (distant-dependent dielectric constant of $4r_{ij}$ with fixed DNA coordinates and all charges reassigned) in preparing the simulation until a derivative of 0.1 kcal•mol^{-1}•Å^{-2} was achieved. All DNA charges were reapplied for minimization protocols. After minimizing the trajectory structures, the sequence affinity, or footprint, was determined by analyzing the graphical representation of the potential energies of the minimized trajectory structures against time.

(iv) Construction of WG-DNA To Enable Sequencing with Side-by-Side or Hairpin Ligands. The core 5′-GGCC-3′ and 5′-ATAT-3′ base pair coordinates of two DNA crystal structures in 2:1 ligand complexes were extracted and taken to represent the groove parameters associated with side-by-side GC and AT-binding lexitropsins respectively (pdb reference 334D[41] and pdb reference 378D[42]). A script was compiled to build the equivalent double stranded tetramers (i.e. 5′-GGCC-3′ and 5′-ATAT-3′), which were systematically changed by translation along the *x* and *y* axis, axial rise, tilt, roll, twist in order to establish the helical parameter for an AT or GC step in WG-DNA in order to build any required sequence to explore. WG-DNA used in subsequent simulations was constructed using the helical parameters shown Table 1. Prior to use, WG-DNA structures were minimized as described above for standardized DNA.

ACKNOWLEDGMENT

We would like to thank Professor David Wemmer for supplying coordinates for some of the models.

Supporting Information Available: Simulation movies. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Fishleigh, R. V.; Fox, K. R.; Khalaf, A. I.; Pitt, A. R.; Scobie, M.; Suckling, C. J.; Urwin, J.; Waigh, R. D.; Young, S. C. DNA Binding, Solubility, and Partitioning Characteristics of Extended Lexitropsins. *J. Med. Chem.* **2000**, *43*, 3257−3266.

(2) Khalaf, A. I.; Pitt, A. R.; Scobie, M.; Suckling, C. J.; Urwin, J.; Waigh, R. D.; Fishleigh, R. V.; Young, S. C. Synthesis of Novel DNA Binding Agents: Indole-Containing Analogues of Bis-Netropsin. *J. Chem. Res.* **2000**, 264−265.

(3) Khalaf, A. I.; Pitt, A. R.; Scobie, M.; Suckling, C. J.; Urwin, J.; Waigh, R. D.; Fishleigh, R. V.; Young, S. C.; Wylie, W. A. The Synthesis of Some Head to Head Linked DNA Minor Groove Binders. *Tetrahedron* **2000**, *56*, 5225−5239.

(4) Anthony, N. G.; Fox, K. R.; Johnston, B. F.; Khalaf, A. I.; Mackay, S. P.; McGroarty, I. S.; Parkinson, J. A.; Skellern, G. G.; Suckling, C. J.; Waigh, R. D. DNA Binding of a Short Lexitropsin. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 1353−1356.

(5) Dudouet, B.; Burnett, B.; Dickinson, L. A.; Wood, M. R.; Melander, C.; Belitsky, J. M.; Edelson, B. S.; Wurtz, N.; Briehn, C.; Dervan, P. B.; Gottesfeld, J. M. Accesibility of Nuclear Chromatin by DNA Binding Polyamides. *Chem. Biol.* **2003**, *10*, 859−867.

(6) Mapp, A.; Ansari, A. Z.; Ptashne, M.; Dervan, P. B. Activation of Gene Expression by Small Molecule Transcription Factors. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 3930−3935.

(7) White, S.; Turner, J. M.; Szewczyk, J. W.; Baird, E. E.; Dervan, P. B. Affinity and Specificity of Multiple Hydroxypyrrole/Pyrrole Ring Pairings for Coded Recognition of DNA. *J. Am. Chem. Soc.* **1999**, *121*, 260−261.

(8) Dickinson, L. A.; Gulizia, R. J.; Trauger, J. W.; Baird, E. E.; Mosier, D. E.; Gottesfeld, J. M.; Dervan, P. B. Inhibition of RNA Polymerase II Transcription in Human Cells by Synthetic DNA-binding Ligands. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*.

(9) Gottesfeld, J. M.; Neely, L.; Trauger, J. W.; Baird, E. E.; Dervan, P. B. Regulation of Gene Expression by Small Molecules. *Nature* **1997**, *387*, 202−205.

(10) Lown, J. W. Design and Development of Sequence Selective Lexitropsin DNA Minor Groove Binders. *Drug Dev. Res.* **1995**, *34*, 145−183.

(11) Neidle, S. DNA Minor-Groove Recognition by Small Molecules. *Nat. Prod. Rep.* **2001**, *18*, 291−309.

(12) Dervan, P. B. Molecular Recognition of DNA by Small Molecules. *Bioorg. Med. Chem.* **2001**, *9*, 2215−2235.

(13) Dervan, P. B.; Edelson, B. S. Recognition of the DNA Minor Groove by Pyrrole-Imidazole Polyamides. *Curr. Opin. Struct. Biol.* **2003**, *13*, 284−289.

(14) Kopka, M. L.; Yoon, C.; Goodsell, D.; Pjura, P.; Dickerson, R. E. The Molecular Origin of DNA-Drug Specificity in Netropsin and Distamycin. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 1376−1380.

(15) Patel, D. J. Antibiotic-DNA Interactions: Intermolecular Nuclear Overhauser Effects in the Netropsin-d(CGCGAATTCGCG) Complex in Solution. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, *79*, 6424−6428.

(16) Wade, W. S.; Mrksich, M.; Dervan, P. B. Design of Peptides that Bind in the Minor Groove of DNA at 5′-(A,T)G(A,T)C(A,T)-3′ Sequences by a Dimeric Side-by-Side Motif. *J. Am. Chem. Soc.* **1992**, *114*, 8783−8794.

(17) Mrksich, M.; Wade, W. S.; Dwyer, T. J.; Geierstanger, B. H.; Wemmer, D. E.; Dervan, P. B. Antiparallel Side-by-Side Dimeric Motif for Sequence-Specific Recognition in the Minor Groove of DNA by the Designed Peptide 1-methylimidazole-2-carboxamide Netropsin. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 7586−7590.

(18) Pelton, J. G.; Wemmer, D. E. Structural Characterization of a 2:1 Distamycin A/d(CGCAAATTGGC) Complex by Two-Dimensional NMR. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 5723−5727.

(19) Kielkopf, C. L.; White, S.; Szewczyk, J. W.; Turner, J. M.; Baird, E. E.; Dervan, P. B.; Rees, D. C. A Structural Basis for Recognition of AT and TA Base Pairs in the Minor Groove of B-DNA. *Science* **1998**, *282*, 111−115.

(20) White, S.; Szewczyk, J. W.; Turner, J. M.; Baird, E. E.; Dervan, P. B. Recognition of the Four Watson−Crick Base Pairs in the DNA Minor Groove by Synthetic Ligands. *Nature* **1998**, *391*, 468−471.

(21) Lane, A. N.; Jenkins, T. C. Thermodynamics of Nucleic Acids and Their Interactions with Ligands. *Quart. Rev. Biophys.* **2000**, *33*, 255−306.

(22) Misra, V. K.; Honig, B. On the Magnitude of the Electrostatic Contribution to Ligand-DNA Interactions. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 4691−4695.

(23) Khalaf, A. I.; Waigh, R. D.; Drummond, A. J.; Pringle, B.; McGroarty, I.; Skellern, G. G.; Suckling, C. J. Distamycin Analogues with Enhanced Lipophilicity: Synthesis and Antimicrobial Activity. *J. Med. Chem.* **2004**, *47*, 2133−2156.

(24) Anthony, N. G.; Johnston, B. F.; Khalaf, A. I.; Mackay, S. P.; Parkinson, J. A.; Suckling, C.J.; Waigh, R. D. Short Lexitropsin that Recognizes the DNA Minor Groove at 5′-ACTAGT-3′: Understanding the Role of Isopropyl-Thiazole. *J. Am. Chem. Soc.* **2004** *126*, 11338-11349.

(25) Laughton, C. A.; Jenkins, T. C.; Fox, K. R.; Neidle, S. Interaction of Berenil with the *tyrT* DNA Sequence Studied by Footprinting and Molecular Modelling. Implications for the Design of Sequence-Specific DNA Recognition Agents. *Nucleic Acids Res.* **1990**, *18*, 4479−4488.

(26) Dolenc, J.; Borštnik, U.; Hodošcek, M.; Koller, J.; Janežic, D. An Ab Initio QM/MM Study of the Conformational Stability of Complexes Formed by Netropsin and DNA. The Importance of van der Waals Interactions and Hydrogen Bonding. *J. Mol. Struct.* **2005**, *718*, 77−85.

(27) Barceló, F.; Capó, D.; Portugal, J. Thermodynamic Characterization of the Multivalent Binding of Chartreusin to DNA. *Nucleic Acids Res.* **2002**, *30*, 4567−4573.

(28) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(29) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(30) Tidor, B.; Irikura, K. K.; Brooks, B.; Karplus, M. Dynamics of DNA Oligomers. *J. Biomol. Struct. Dyn.* **1983**, *1*, 231−252.

(31) Singh, U. C.; Weiner, S. J.; Kollman, P. Molecular Dynamics Simulations of d(C−G−C−G−A) X d(T−C−G−C−G) With and Without “Hydrated” Counterions. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 755−759.

(32) Levitt, M. Computer Simulation of DNA Double-Helix Dynamics. *Symposium of Quantic Biology*; Cold Spring Harbour: 1983; pp 251−262.

(33) Behling, R. W.; Rao, S. N.; Kollman, P.; Kearns, D. R. Molecular Mechanics and Dynamics Calculations on (dA)10.(dT)10 Incorporating Distance Constraints Derived from NMR Relaxation Measurements. *Biochemistry* **1987**, *26*, 4674−4681.

(34) Ravishanker, G.; Swaminathan, S.; Beveridge, D. L.; Lavery, R.; Sklenar, H. Conformational and Helicoidal Analysis of 30 ps of Molecular Dynamics on the d(CGCGAATTCGCG) Double Helix: “Curves”, Dials and Windows. *J. Biomol. Struct. Dyn.* **1989**, *6*, 669−699.

(35) Boehncke, K.; Nonella, M.; Schulten, K. Molecular Dynamics Investigation of the Interaction Between DNA and Distamycin. *Biochemistry* **1991**, *30*, 5465−5475.

(36) Bates, P. J.; Laughton, C. A.; Jenkins, T. C.; Capaldi, D. C.; Roselt, D.; Reese, C. B.; Neidle, S. Efficient Triple Helix Formation by Oligodeoxyribonucleotides Containing Alpha- or Beta-2-amino-5-(2-deoxy-D-ribofuranosyl) Pyridine Residues. *Nucleic Acids Res.* **1996**, *24*, 4176−4184.

(37) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. Comparison of Automated Docking Programs as Virtual Screening Tools. *J. Med. Chem.* **2005**, *48*, 962−976.

(38) Reddy, S. Y.; Leclerc, F.; Karplus, M. DNA Polymorphism: A Comparison of Force Fields for Nucleic Acids. *Biophys. J.* **2003**, *84*, 1421−1449.

(39) Foloppe, N.; MacKerell, A. D. Intrinsic Conformational Properties of Deoxyribonucleosides: Implicated Role for Cytosine in the Equilibrium Among the A, B and Z forms of DNA. *Biophys. J.* **1999**, *76*, 3206−3218.

(40) Abu-Daya, A.; Brown, P. M.; Fox, K. R. DNA Sequence Preferences of Several AT-Selective Minor Groove Binding Ligands. *Nucleic Acids Res.* **1995**, *23*, 3385−3392.

(41) Kopka, M. L.; Goodsell, D.; Han, G. W.; Chiu, T. K.; Lown, J. W.; Dickerson, R. E. Defining GC−Specificity in the Minor Groove: Side-by-Side Binding of the Di-Imidazole Lexitropsin to C-A-T-G-G-C-C-A-T-G. *Structure* **1997**, *5*, 1033−1046.

(42) Mitra, S. N.; Wahl, M. C.; Sundaralingam, M. Structure of the Side-by-Side Binding of Distamycin to d(GTATATAC)(2). *Acta Crystallogr. D* **1999**, *55*, 602−609.

(43) Parks, M. E.; Baird, E. E.; Dervan, P. B. Optimization of the Hairpin Polyamide Design for Recognition of the Minor Groove of DNA. *J. Am. Chem. Soc.* **1996**, *118*, 6147−6152.

(44) Winston, C. T.; Takahiro, I.; Dale, L. B. Comprehensive High-Resolution Analysis of Hairpin Polyamides Utilizing a Fluorescent Intercalator Displacement (FID) Assay. *Bioorg. Med. Chem.* **2003**, *11*, 4479−4486.

(45) Swalley, S. E.; Baird, E. E.; Dervan, P. B. Discrimination of 5′-GGGG-3′, 5′-GCGC-3′, and 5′-GGCC-3′ Sequences in the Minor Groove of DNA by Eight-Ring Hairpin Polyamides. *J. Am. Chem. Soc.* **1997**, *119*, 6953−6961.

(46) James, P. L.; Merkina, E. E.; Khalaf, A. I.; Suckling, C. J.; Waigh, R. D.; Brown, T. Fox, K. R. DNA Sequence Recognition by an Isopropyl

Substituted Thiazole polyamide. *Nucleic Acids Res.* **2004**, *32*, 3410−3417.

(47) Discover 2.9.5/94.0 User Guide: San Diego, CA, 1994.

(48) Stewart, J. P. P. *MOPAC 6.0*; available from the Quantum Chemistry Programme Exchange, Indiana University, Bloomington, IN.

(49) de Oliveira, A. M.; Custódio, F. B.; Donnici, C. L.; Montanari, C. A. QSAR and Molecular Modelling Studies on B-DNA Recognition of Minor Groove Binders. *Eur. J. Med. Chem.* **2003**, *38*, 141−155.

CI050153B