Methods 47 (2009) 159–167

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Review Article

Thermogenomics: Thermodynamic-based approaches to genomic analyses of DNA structure

P. Shing Ho *

Department of Biochemistry and Molecular Biology, 1870 Campus Delivery, 316 MRB Building, Colorado State University, Fort Collins, CO 80523-1870, USA

ARTICLE INFO

Article history:
Accepted 12 September 2008
Available online 10 October 2008

Keywords:
Z-DNA
Nuclear factor I
Transcription regulation

ABSTRACT

The postgenomic era is all about learning about function by comparing genomic sequences within and between organisms. This review describes an approach that applies detailed thermodynamic information, as opposed to sequence motif searches, to analyze genomes (thermogenomics) for the occurrence of sequences with the potential to form left-handed Z-DNA and those that bind the eukaryotic nuclear factor I (NFI) transcriptional regulators. Such thermogenomic strategies allow us to address the questions of whether Z-DNA forming sequences can potentially function in regulating transcription of eukaryotic genes and how such function may emerge relative to other GC-rich elements, such as NFI recognition sites, to become a transcriptional coactivator.

1. DNA structures and their functions

1.1. Summary of DNA structures

It is now over 55 years since James Watson and Francis Crick [1] proposed a right-handed double-helical structure of DNA (the B-DNA structure) as a means to replicate and express the genetic information in a cell. Since that landmark paper, DNA has proven to be a highly polymorphic biomolecule, with a number of alternative structures identified, many of which are associated with specific cellular functions. For example, the four-stranded G-quartets are found at telomer ends of chromosomes and are associated with cellular senescence [2,3] and cancer progression [4,5]. However, the assumption is that the original B-form of DNA is the dominant conformation in the cell; indeed, for most biologists, all that is required to understand the biological function of DNA is the generic antiparallel Watson–Crick base paired double-helix.

Still, the plethora of non-B-DNA conformations has become exceedingly large—indeed, the historical approach of naming new DNA structures using letters of the alphabet has run into the problem that there are few letters from A to Z [6] that have not already been assigned, leaving very few letters available to name new conformations (Fig. 1). Table 1 provides a summary of some structurally interesting conformations, but is not meant to be exhaustive as there have been a number of reviews on all the various DNA structure [7,8]. Nearly all of these conformations have been characterized by X-ray diffraction studies on fibers and/or crystals, or by NMR; thus, there is plenty of detailed structural information, often times to atomic resolution, on these various DNA conformations. There is, however, much less known about how or if any of these structures contribute to genetic functions in the cell.

A very interesting case in point is Z-DNA, a rigid, left-handed variant of the standard right-handed double-helix [9]. Although it was the very first DNA structure to be crystallographically determined at atomic resolution (the atomic structure of B-DNA followed a couple of years later [10]), its potential biological function has been debated for the thirty years since its discovery [11–14]. The initial strategies to establish a biological function for Z-DNA focused primarily on biochemical and biophysical studies, including mapping its occurrence in functional chromosomes with Z-DNA specific antibodies and chemical reagents, along with attempts to isolate and identify functional proteins with specificity for the left-handed double-helix. It had been the missteps and inability during these early concerted efforts to identify “Z-proteins” that made the structure a near pariah in biology [14]. In recent years, a number of proteins have in fact been found to recognize the structure of Z-DNA in a sequence-independent manner, including proteins associated with RNA editing [15], gene transactivation [16], etc. However, it is also known that Z-DNA does not bind to proteins that either require the base-sequence information in the grooves of the right-handed helices for recognition (including, for example, transcription factors), or require the helical flexibility of B-type DNAs, such as the histone proteins that package DNA into nucleosome structures [17,18]. This latter property, in fact, has been proposed as providing a means for Z-DNA to serve as a transcriptional coactivator [19]. This review will focus on the location of Z-DNA in genomes as an example of how genomic

Abbreviations: ZDR, potential Z-DNA regions; NFI, nuclear factor I; CSF, colony stimulating factor.

* Fax: +1 970 491 0494.
E-mail address: Shing.Ho@colostate.edu.

1046-2023/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2008.09.007

and phylogenomic analyses can help to elucidate the functional role of a structure as well as its evolutionary emergence relative to other functional elements in the cell.

In 1985, before genomics and bioinformatics were catch phrases, a few postdoctoral associates in the laboratory of Dr. Alexander Rich decided that we could take a computational approach to addressing the question of whether the left-handed structure had any biologically interesting function. To do this, we developed a computer program (called ZHUNT) to study the occurrence of sequences with high propensities to form Z-DNA in genomes [20]. The idea was that if the structure has a function, then sequences with the propensity to adopt this left-handed form would localize, accumulate, or be suppressed in genomic domains with known functions. If, for example, Z-DNA were to be involved in transcriptional regulation, it should be found near markers that control gene expression, including the transcription start-site (TSS) [21], or be coupled with binding sites for transcription factors. Such a bioinformatics approach should be general and not depend on finding any specific protein that recognizes the structure. This review will discuss *in silico*, thermodynamics based approaches to identifying and locating non-B-DNA structures in genomes.

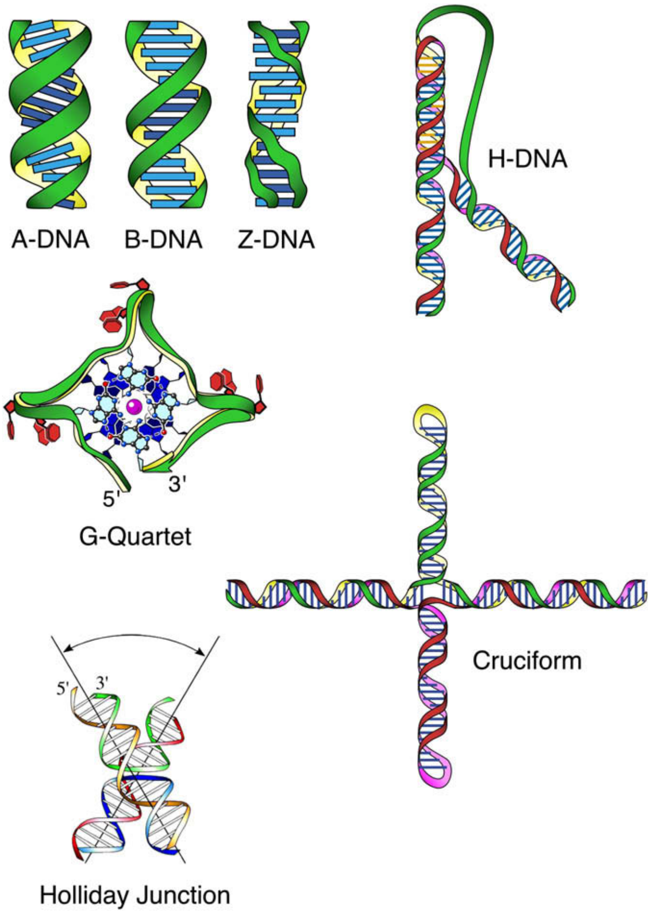

Fig. 1. Structures of various DNA conformations. Shown are the double-stranded conformations of A-, B-, and Z-DNA (ribbons trace the phosphodeoxyribose backbone, light bars represent the base pair faces in the major groove, dark bars for the minor groove), the triple-stranded H-DNA form, the G-quartet structure of the human telomer sequence [66], the extruded cruciform structure, and the Holliday junction [61] at the core of cruciforms and other recombination intermediates.

Table 1
Description of several DNA forms and their sequence dependences

DNA form	Description/potential function	Sequence dependence
A-DNA	Right-handed duplex with Watson–Crick base pairs, 11 bp/turn, 2.6 Å helical rise Function: Implicated in RNA polymerase recognition [56]	Nonalternating GC-rich; GGN, NGG and CC(C/G)
B-DNA	Right-handed duplex with Watson–Crick base pairs, 10–10.5 bp/turn, 3.4 Å rise Function: Canonical structure of DNA	All sequences
Cruciforms	Extruded DNA duplexes with 2 B-DNA stem-loops connected by a four-way junction (see Holliday junction)	AT-rich inverted repeats
"Extended"—DNA	Extended right-handed A-like duplex with Watson–Crick base pairs; 12.2 bp/turn, 3.6–3.7 Å rise Function: Intermediate in B- to A-DNA pathway; spontaneous deamination of mC in transition mutation to T [49]	Cytosine methylation in GC-rich sequences
G-Quartet	Four-stranded structure with G·G·G·G Hoogsteen type base pairs Function: Telomeric ends [57–59]	Stings of consecutive G's
H-DNA	Three-stranded helix with Watson–Crick and Hoogsteen type base triplets Function: Mutagenesis in mammalian cells [60]	Mirror repeats
Holliday Junctions	Four-stranded structure with B-type arms Function: Recombination and recombination dependent processes (DNA repair, etc.) [61]	GC-rich inverted repeats: NYC (N = A > G > C; Y = C > T)
i-Motif	Four-stranded structure with intercalated C·C+ base pairs Function: Telomer ends of human chromosomes [62–64]	Strings of C's
Z-DNA	Left-handed duplex with Watson–Crick base pairs, −12.0 bp/turn, 3.7 Å rise Function: RNA editing; transcription regulation; DNA deletion [12,65]	Alternating YR dinucleotides (Y = C > T; R = G > A)

Listed are a series of DNA structures, brief description of their structural features and potential functions, and their sequence dependency for formation.

There are now many widely available tools to search for simple sequence motifs [22] and, since all DNA conformations have sequence preferences for their formation, one might expect that searching for unusual conformations would be relatively straightforward since most non-B-DNA type structures are typically characterized by simple repeating motifs. Inverted and mirror repeats, for example, are motifs that are characteristic of cruciforms and H-DNA and, therefore, algorithms that search for such sequence patterns can be used to identify the genomic occurrences of and potential functions for such structures [23–25]. Using, again, Z-DNA as the example, our experience is that simple sequence rules do not accurately identify the location or, more importantly, the probability or propensity of a sequence to adopt the left-handed conformation. For example, if we were to search for one turn of the two sequences that are known to form Z-DNA (CGCGCGCGCGCG, and CACACACACACA), statistically, we would expect to find just 6 CpG type and 60 of the CpA type sequences in 1 billion base pairs (for the human genome, which is 41.5% GC in content). We know, however, that such sequences are much more prevalent than predicted. In addition, many eukaryotic promoters are not strictly simple repeats of CpG or CpA/TpG dinucleotides (for example, promoter sequences of the rat prolactin gene [26]) and, therefore, such sequences would not be identified in a search for repeating elements, nor would we have a measure of their propensities to adopt the left-handed form relative to these standard repeating elements. Thus, this review will focus primarily on approaches that apply thermodynamic rules to search for DNA structures, particularly

Z-DNA, and other functional elements, including sequence motifs that are recognized by the eukaryotic nuclear factor I (NFI) transcriptional element. These particular cases will provide useful insights into methods for the development and application of thermodynamic genomics (or "thermogenomics") that would be generally applicable to studying other alternative DNA structures as well as functional elements that are based on B-type DNAs. I will also discuss how phylogenomic analysis of Z-DNA along with other elements provides a model for how various related transcriptional elements emerged and evolved across the genomes of various organisms.

2. Genomic searches for DNA structures

The simplest approach to finding DNA conformations is to look for sequences or sequence motifs that characterize that structure. The starting assumption is always that the typical sequence will adopt the canonical B-DNA double-helix under standard or physiological conditions. Nearly all non-B-structures, however, are characterized not by a single simple sequence or set of sequences, but by sequence motifs [27] that are have various propensities to undergo structural transitions under conditions that deviate from the standard (Table 1).

This review will start with a detailed description of how thermogenomic methods were developed to search for sequences with high propensities to form left-handed Z-DNA or to bind eukaryotic nuclear factor I. This will be followed by discussion of specific examples of genomic and phylogenomic analyses that apply these thermogenomic algorithms, and what can be learned from such studies.

2.1. Thermogenomic search for Z-DNA

2.1.1. The structural and thermodynamic properties of Z-DNA

It was recognized from the initial crystal structure that Z-DNA is characterized by a peculiar structural alternation of nucleotides with the bases in their normal *anti* form (where the base extends away from the deoxyribose backbone, as in B-DNA) and in the less stable *syn* form (where the base rotates and stacks over the deoxyrobose ring)—it is this alternating conformation of the base that generates the zig–zag pattern of the Z-DNA backbone (hence its name). Z-DNA is characterized as having a dinucleotide repeat, because of this strict alternating pattern of conformations, as compared to the nucleotide repeat of other DNA duplex structures. Since pyrimidine nucleotides do not adopt the unfavorable *syn* form, this alternating pattern of *anti-syn* conformations accounts for the preference for repeating alternating pyrimidine-purine (APP) motifs, with CpG > CpA/TpG > TpA dinucleotides in terms of their propensity to form Z-DNA. However, as we will see later on, a simple search for repeating strings of dinucleotides will miss some very important, potentially functional Z-DNA sequences, and this is the reason for taking a thermogenomics approach.

2.1.2. The B–Z transition and the two-state zipper model

As recently as 2006, there had been significant debate as to how right-handed B-DNA duplexes can convert to a left-handed form, with over a dozen different models published that provide a molecular and/or thermodynamic description of the transition [28]. The debate has recently been quelled when the single-crystal structure of the junction that splices the right- and left-handed duplexes (the B–Z junction) was determined [29] (Fig. 2). The structure shows that the junction involves the breaking and opening of one base pair such that the bases are protruding out and away from the DNA duplex. The segments that remain as intact duplexes with opposite hands have their helical axes staggered in such a way as to maintain the stacking of the bases at the termini of each duplex segment. Consequently, the phosphoribose backbones of these helical segments point in opposite directions (the so-called "chain sense paradox"). To accommodate the stagger in the backbones, the phosphoriboses of the B–Z junction itself are oriented nearly perpendicular to both the right- and left-handed helical domains. Thus, although there is a single base pair that is formally melted, the backbone associated with terminal base pairs of each helical segment is also distorted and lies relatively exposed. This would account for the approximate 4 base pairs that are typically seen to be sensitive to single-strand specific nucleases [30] and chemicals [31]. The molecular model, therefore, is consistent with the features of the B–Z junction seen in solution.

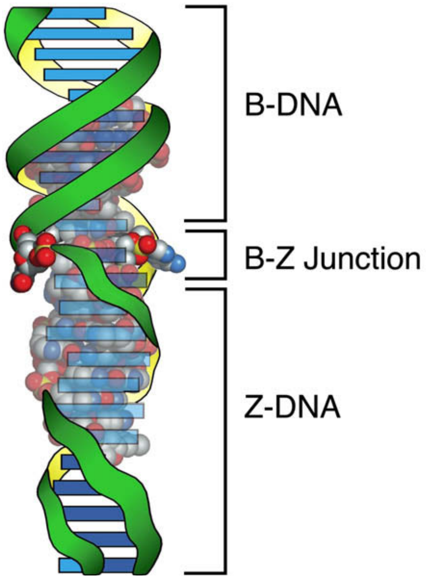

Fig. 2. Structure of the junction that abuts right-handed B-DNA to left-handed Z-DNA (the B–Z junction). The atoms at the junction are shown as van der Waals spheres [29], while the backbones of the adjoining B- and Z-DNA duplexes are traced by ribbons.

Such a detailed molecular model, however, is not essential when trying to predict the occurrence of Z-DNA in genomes. What is necessary is a model that describes the thermodynamic propensities for different sequences to convert from right-handed B-DNA to left-handed Z-DNA (Fig. 3). Such a model is the all-or-none zipper model that is analogous to the helix-coil model for the melting of helical structures in proteins. The two-state zipper model for the B–Z transition assumes that any base pair in a sequence can adopt either the B- or Z-forms. If we start with all B-DNA in a sequence, the B–Z transition is initiated with the formation of 2 B–Z junctions (this is the high-energy, low probability nucleation step). Once nucleated, migration of the two junctions in opposite directions allows the formation of Z-DNA between them (this is the propagation step.

It is important to note at this point that Z-DNA can be induced by negative superhelical stress [32,33]. If, for example, an alternating CpG sequence (e.g., 20 dinucleotides of CpG, or $(CpG)_{20}$) is placed in a closed circular plasmid, and that plasmid is negatively supercoiled (with negative linking numbers, ΔLk), then, at some point, the amount of energy released by converting that sequence from right-handed to left-handed Z-DNA (change in helical twist, or ΔTw) will relax sufficient numbers of supercoils (change in writhe, or $\Delta Wr = \Delta Lk - \Delta Tw$) to overcome the energy required to induce a B–Z transition ($\Delta G^o_{\Delta Wr} = K\Delta Wr^2 = \Delta G^o_{B-Z}$, where $\Delta G^o_{\Delta Wr}$ is the free energy released in terms of the number of supercoils relaxed, ΔG^o_{B-Z} is the total free energy required to induce the B–Z

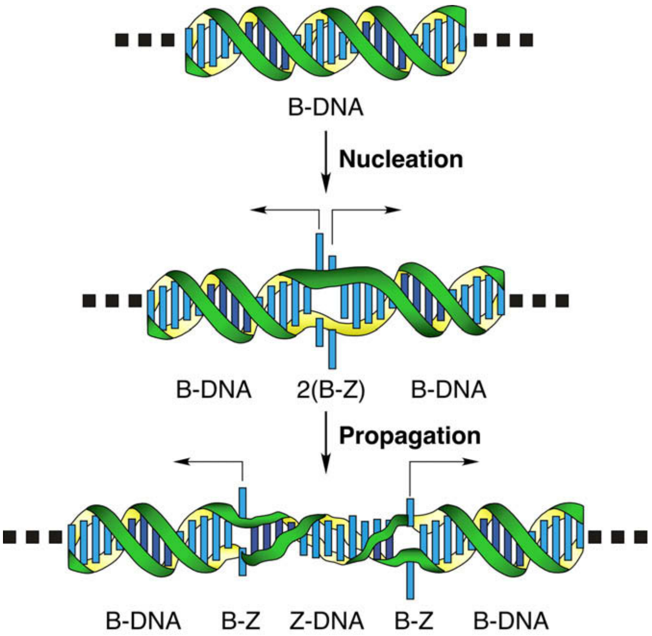

Fig. 3. The two-state zipper model for the transition from B- to Z-DNA (the B–Z transition). From the starting B-DNA duplex, the transition initiates with the formation of 2 B–Z junctions (Fig. 2), which migrate in opposite directions to propagate left-handed DNAs between the junctions.

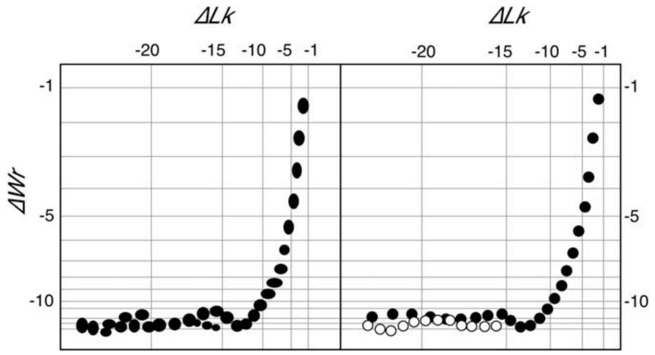

Fig. 4. Two-dimensional gel electrophoresis analysis of a topoisomerase ladder and simulation of the results using ZHUNT. (A) The topoisomer distribution of closed circular DNAs (ccDNAs) at different linking numbers (ΔLk) relative to relaxed ccDNAs ($\Delta Lk = 0$) are resolved along the horizontal direction, while the superhelical twist (writhe, ΔWr) relative to relaxed ccDNA as B-DNA are resolved along the vertical direction. The discontinuity along the vertical direction is indicative of the unwinding of the DNA duplex (twist, ΔTw) relative to standard B-DNA, which is indicative of formation of a structure such as left-handed Z-DNA (for example, the topoisomer at $\Delta Lk = -15.5$ turns has a $\Delta Wr \approx -11.0$ turns, resulting in a $\Delta Tw = \Delta Lk - \Delta Wr = -4.5$ turns). The results are for a sequence found in the promoter region of the rat prolactin gene [26]. (B) Simulation of the gel in A using the ZHUNT algorithm and energy terms in [35] shows that the transitions seen in the rat prolactin promoter is very likely undergoing a transition to classic Z-DNA, as opposed to an alternative left-handed conformation [41].

transition in that sequence, and K is a proportionality constant analogous with the spring constant for the supercoiled DNA).

A statistical mechanics treatment of this two-state zipper model for the B–Z transition was first derived by Peck and Wang in 1983 for the transition in $(CG)_n$ sequences [34], and extended to CA/TG dinucleotides by Vologodskii and Frank-Kamenetskii in 1984 [35]. The experimental values for the steps in the B–Z transition were determined experimentally by measuring the change in topological properties of a series of closed circular plasmid DNAs with increasing superhelical twist (a topoisomer ladder) through two-dimensional gel electrophoresis. In these gels, the topoisomerase ladder is resolved according to the change in superhelical writhe (ΔWr) in the first dimension and the change in linking number (ΔLk) in the second (Fig. 4), taking standard relaxed B-DNA as the reference structure of the closed circular plasmid. The discontinuity in the ΔLk and ΔWr is taken as the change in the helical twist (ΔTw) associated with the B–Z transition, which can be plotted as a function of change in ΔLk to determine the degree of cooperativity and the overall free energy (ΔG°) associated with the transition, determined as follows

$$\Delta G^\circ = K(\Delta Lk_m - \Delta Tw_m)^2 \tag{1}$$

where ΔLk_m and ΔTw_m are the linking number and twist at half the maximum change in twist, and K is taken as $1100RT/N$, for N number of base pairs in the DNA plasmid).

Fitting the statistical mechanics treatment of the zipper model to the two-dimensional gel electrophoresis results provides estimates for the nucleation (σ) and propagation terms (S) of the zipper model. The nucleation term is associated with the formation of the B–Z junction, which is relatively sequence-independent and has an energy requirement of 5 kcal/mol/junction (or 10 kcal/mol for the two junctions for a Z-DNA segment). Thus,

$$\sigma = \mathrm{e}^{-(10\ \mathrm{kcal/mol}+K[\Delta Lk-(-0.8)]^2)/RT} \tag{2}$$

where -0.8 is the unwinding associated with the formation of the B–Z junctions. The sequence dependency for the formation, therefore, is seen in the energy of the propagation step. The propagation energies were determined by Peck and Wang [34] for alternating CpG dinucleotides and by Vologodskii and Frank-Kamenestskii [36,37] for CpA/TpG dinucleotides. Ellison et al. subsequently provided propagation energies for TpA dinucleotides [38] as well as the non-alternating CpC/GpG [20] and CpT/ApG [39] dinucleotides. Using this set of experimental results, a complete set of energies for the propagation of Z-DNA was derived for all dinucleotide combinations [35] that can be applied to any sequence of DNA. Thus,

$$S_j = \mathrm{e}^{-\left(\Delta G^o_{\mathrm{B-Z}}+K[\Delta Lk-j\{-0.36\mathrm{turns/dn}\}]^2\right)/RT} \tag{3}$$

where $\Delta G^\circ{}_{\mathrm{B-Z}}$ is the propagation free energy for a given dinucleotide, j is the number of dinucleotides propagated, and -0.36 is the change in twist associated with the formation of Z-DNA in a dinucleotide.

With these values, the energies are related to the topological changes in the B–Z transition in a supercoiled system through the zipper model by first defining a partition function (Q) as

$$Q = 1 + \sum_{i=1}^{n}\sum_{k=1}^{n}\sigma\left(\prod_{j=i}^{k}S_j\right)\exp\left\{\frac{-K}{RT}\left(\Delta Lk - \left[\sum_{j=i}^{k}0.36j\right] - 0.8\right)^2\right\} \tag{4}$$

for propagation of Z-DNA though n number of dinucleotides, and 0.179 is the ΔTw for each dinucleotide and 0.4 is the ΔTw of the two B–Z junctions. The amount of Z-DNA can thus be predicted as the change in overall twist of the plasmid at any ΔLk according to:

$$< \Delta Tw > = Q^{-1}\left[\sum_{i=1}^{n}\sum_{k=1}^{n}\left(\left[\sum_{j=i}^{k}0.36j\right] - 0.8\right)\sigma\left(\prod_{j=i}^{k}S_j\right) \times \exp\left\{\frac{-K}{RT}\left(\Delta Lk - \left[\sum_{j=i}^{k}0.36j\right] - 0.8\right)^2\right\}\right] \tag{5}$$

2.1.3. ZHUNT: A program for predicting Z-DNA propensities

ZHUNT was constructed as a program to predict the formation of Z-DNA for any sequence of n dinucleotides. A branching algorithm is used to find the minimum total propagation energy for the sequence to assign the ΔG° values to calculate the propagation term (S) for each dinucleotide in the sequence. The program then applies these S terms in a simple set of nested DO LOOPS for the product and summation terms of Q and <ΔTw> as the program walks through a sequence. The complete output from the program includes the "best" stretch of Z-DNA dinucleotides in the sequence, the *anti-syn* assignments for the dinucleotides in this stretch, the <ΔTw> and ΔLk_m for the stretch—the lower the ΔLk_m, the higher the potential that the sequence will form Z-DNA.

In order for the ΔLk_m values to be useful as a predictive tool, they are converted to propensities for forming Z-DNA relative to random sequences. Thus, the ΔLk_m value is compared to those for a set of randomly generated sequences to calculate a propensity (originally called a Z-score [20], but now referred to as P_Z [40]), which reflects the propensity for the sequence to form Z-DNA. P_Z for a particular sequence is defined as the number of random sequences that one must search in order to find one that has a similar or higher propensity to form Z-DNA, and has units of base pairs (bp).

2.1.4. Validation of ZHUNT

As with any computational approach to search for a structural domain, one must ask how accurately does the program locate and predict the propensity for that structure along a sequence. The initial description for the development of the ZHUNT algorithm compared the program predictions to experimental data derived from anti-Z-DNA antibody binding to, for example, the double-stranded replicative form of the viral ϕX-174 genome. In this comparison, the P_Z values from ZHUNT correlated extremely well with the positions and relative binding of anti-Z-DNA antibodies along the viral sequence [20].

A more complete validation of the basic statistical mechanics algorithm implemented in ZHUNT came in 1994 when the program was used to address the following question: Do CA/TG repeats in fact adopt the Z-DNA structure, as opposed to, for example, extruded cruciforms? This latter question was raised in an intriguing paper by Kladde et al. [26]. In this study, the authors had identified a set of CA/TG-type sequences at the 5′-promoter of the rat prolactin gene that, when placed under negative superhelical stress, become unwound. A detailed analysis of the degree of unwinding showed the resulting non-B structure to have a helical twist of −13 to −21 base pairs per turn (negative because it is left-handed), as compared to −12.0 for the standard Z-DNA model. In addition, the chemical footprinting of the "unpaired" nucleotides in CA/TG sequences under negative superhelical stress were shown to not reside exclusive at the edges of the sequence, but distributed well within the sequence itself. The conclusion was that CpA/TpG type sequences and promoters formed some type of left-handed structure that was not quite as left-handed as Z-DNA.

An analysis of the sequence using ZHUNT, however, indicated that the properties observed were entirely consistent with the CpA/TpG type sequences forming standard Z-DNA [41]. Using the energies and the degree of unwinding associated with Z-DNA, the behavior of these types of sequences could be entirely simulated, including the complex migration patterns, the patterns of nuclease digestion to follow the progression of the B–Z transition and location of junctions, the degree of unwinding. Basically, the anomalous behavior of CpA/TpG types sequences could be simply attributed to the less cooperative transition in such sequences (as predicted by the higher propagation energies) relative to the standard alternating CpG sequences normally associated with Z-DNA. Thus, the zipper model described here very accurately predicts and simulates the overall and detailed behavior of sequences that can and do form left-handed Z-DNA, and, therefore, they can be used confidently to identify regions of genomes with the potential to form Z-DNA (which we now call ZDRs [40]).

ZHUNT is most conveniently accessed on-line at http://gac-web.cgrb.oregonstate.edu/zDNA/index. The limitation is that this is hosted on a server at Oregon State University, which has set the limit of the analysis to 1 megabase pair. However, for simple sequences (e.g., a gene), this is the most convenient way to use the program. Simply upload a sequence in FASTA format onto your local computer, then upload the file to the ZHUNT server and submit. The server will provide a job number, which allows you to retrieve the results later. The output will provide an analysis of the nucleotide composition of the sequence, and the starting nucleotide position, the length, the propensity, and the sequence of any ZDR it finds (with P_Z set as $\geqslant$700 bp).

For larger sequences and more complete analyses, you will need the full ZHUNT program, which can be obtained from the author. After agreeing to the conditions, you will be provided with the source code for the program to compile on your own computer. The output from the full program is a text file list, starting at nucleotide 1, of the P_Z, length, ΔLk_m, sequence, and conformation assignments for the dinucleotides for the best ZDR starting at that nucleotide. For genomic sequences, this can be a very large file and, therefore, you should insure that there is sufficient disk space to handle the output.

2.2. Thermogenomic approach to analysis of nuclear factor I (NFI) binding sites

A different approach to studying DNA structure and function based on thermodynamic behavior is to search for recognition sites for DNA binding proteins based on the free energies of binding, or affinity constants. A good example of how thermogenomics works for this type of question can be seen with the DNA binding sites for nuclear factor I (NFI) [42], which recognizes a eukaryotic CAAT-box transcription promoter sequence [43–45]. The promoter, with a consensus sequence TTGGCN_5GCCAA, where N_5 is a spacer of any 5 nucleotides, is 60% GC and, therefore is considered to be a GC-rich promoter. Our interest in developing a thermogenomic approach to identifying NFI sites came, again, from the model of Liu et al. [19], for coactivation in the human CSF-1 gene, with the immediate question being: How often are ZDRs coupled with NFI sites?

The key to developing an algorithm for NFI sites came from the studies of Roulet et al. [46] in which the authors systematically determined the effect of nucleotide substitutions at each position of the promoter sequence on NFI binding. From this work, a set of binding scores were derived to describe the relative affinities of NFI for each sequence variant.

In order to define a measure of NFI binding propensities that is comparable to P_Z [40], the binding scores of Roulet et al. [46] were applied to a large set of random sequence, an average binding score for the population calculated, and the deviation from this average defined. From this, we can calculate P_{NFI}, which is defined statistically for a given sequence as the number of random sequences that must be searched to find one that is as good or better at binding NFI. Interestingly, the resulting P_{NFI} mirror the affinity constants for the various sequences, except that the propensities are ~2000 times smaller than the binding constants. In other words, a sequence that is uniquely found in 1000 random nucleotides will have a $K_A \approx 2 \times 10^6$ (or, equivalently, an approximate 0.5 μM dissociation constant). Alternatively, one would expect to find at least one sequence with sub-nanomolar dissociation once every 500 kb random sequences. We expect that the uniqueness of a recognition sequence will always be tied to the affinity of its associated DNA binding protein in similar fashion.

We can apply this NFI analysis to a particular gene to see what can be learned about potentially functional sites. The human colony stimulation factor 1 (CSF-1) gene (plus 1 kb upstream) is

~21 kb in length, suggesting that there should be fewer than 1 site with subnanomolar affinity for NFI, according to the simple correlation between uniqueness and affinity described above. The analysis of this gene shows greater than six such sites, indicating that this eukaryotic gene is not random [40], but accumulates such sequences (Fig. 5). The question, then, is how does the transcriptional machinery know which of the tight binding sites would be functional. Obviously, there are other signals for transcriptional regulation, but, in this case, we see that ZHUNT located only 2 ZDRs in the sequence, and only one sits near the TSS, and that is just upstream of an NFI site. We can, therefore, conclude that NFI sites are not randomly distributed across a genome, and that even high affinity sites need additional genomic markers to distinguish them as being functional. This leads to the broader approach of genomic analyses for these and other potentially functional elements in the DNA.

3. Genomic analyses

A common question that is asked in an analysis of DNA structure is: "Where is that structure found in a gene, in particular, the specific gene sequence that I am studying?" More often, the question is raised as "Hmm, I see a repetitive sequence—I wonder if it is interesting or important?" If it fits into one of the simple repeating motifs in Table 1, then it is tempting to speculate on the involvement of a non-B-DNA structure. Liu et al. [19], for example, observed a repeating pattern of alternating CA/TG dinucleotides just upstream of a classic CAAT promoter sequence in the human CSF-1 gene. Through a systematic series of studies in which the repeating pattern was scrambled, the sequence was shifted downstream, and the sequence probed experimentally for a structural transition, it was concluded that this sequence was an important component for the activation of the gene and that this activation involved a transition to left-handed Z-DNA. A model was proposed in which Z-DNA served as a coactivator, along with NFI binding to the CAAT-box promoter, by preventing nucleosomes from reassociating on the regulatory elements of the activated CSF-1 gene.

This well controlled study raised once again the potential involvement of Z-DNA in the regulation of gene expression. Does this mean, however, that this is a common regulatory mechanism, or is it specific just for this one gene or family of genes? Do we need to go through such exhaustive functional mapping of each repeating CA/TG motif in the genome in order to invoke a regulatory function to Z-DNA? If we are looking for a general role for a structure like Z-DNA, then the answer is "No". We can ask where does the structure localize in a genome, and is there a pattern of accumulation or suppression at the various known functional markers of genes.

To tackle such questions, we must first ask what constitutes a reasonable propensity (P_Z or P_{NFI} for Z-DNA and NFI binding, respectively) to use for identifying potential sequences of interest. If we start with the assumption that both ZDRs and NFI sites are functional in the regulation of some human genes, as suggested by the studies of Liu et al. [19], then we would expect that such sequences should occur, at some level of P_Z and P_{NFI}, at a higher probability than in random sequences. The number of ZDRs identified in human chromosome 22 starts to deviate significantly from random sequences with the same C + G content as the chromosome at $P_Z \geqslant 700$ bp [40]. This is equivalent to a 12 bp stretch (or one full turn) of alternating CpA/TpG dinucleotides, which has been shown experimentally to form Z-DNA in negatively supercoiled closed circular plasmids [31]; thus, this statistical approach to identifying ZDRs mirrors what we would expect to be a good Z-DNA forming sequence from experiment. A similar analysis indicates that $P_{\mathrm{NFI}} \geqslant 500$ bp are nonrandom, while $P_{\mathrm{NFI}} \geqslant 5000$ bp were considered to be strong and potentially functional [40], distinctions that are consistent with the ability of these sequence classes to be recognized by NFI and, subsequently, to enhance expression of a reporter gene [46]. With these definitions, sequences with a potential to form Z-DNA or bind NFI can be identified in a genome in a quantitative manner based on accurate thermodynamic and statistical criteria.

The key to a genome-wide search for DNA structure is to know what specific question is being asked. If the question is "Do ZDRs play a role in regulating transcription in a particular organism?", then, we can apply a thermogenomic approach to determine the:

1. General occurrence of ZDRs in the genome,
2. Distribution of ZDRs across the genome relative to the distribution of genes and predicted genes, and
3. Distribution of ZDRs across functional markers of genes.

For prokaryotic systems, we can address the first question of the general occurrence of Z-DNA in the genome. The number and location of genes in prokaryotic systems are very well determined. There were a large number of ZDRs identified in the *Escherichia coli* genome, but their pattern of occurrence was essentially random across the sequence [40]. In addition, ZDRs are seen to be suppressed just upstream of the transcription start-sites (TSS) and at the 3′-end of genes [21,40,47]. This reflects the very well defined sequences of the promoters and stop codons in prokaryotes.

In eukaryotes, the location of genes, pseudogenes, etc. are not as well defined. Therefore, it is very important in such analyses that the genomes are well annotated. For our studies, we define genes

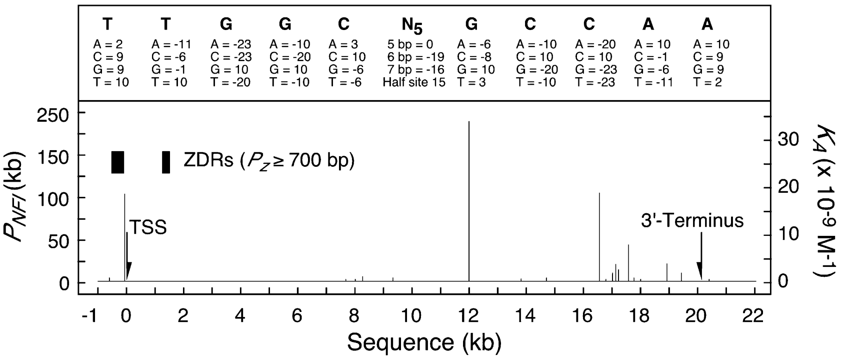

Fig. 5. Distribution of NFI binding sites and ZDRs along the human CSF-1 gene. The top panel shows the consensus sequence and the effect of each perturbation in the sequence on the binding scores, as defined by Roulet et al. [46]. The bottom panel shows the high propensity NFI binding sites along with their P_{NFI} values and association binding constants (K_A). ZDRs identified in the sequence by ZHUNT are shown as bars [40].

as those that are confirmed by cDNA analysis, and classify sequences from GENSCAN [48] to be predicted genes. Even for genes that are confirmed by cDNAs, the location of the actual TSS may not be well defined because of the ambiguities in identifying functional promoter sites. Given these caveats, it was seen that the distributions of both NFI sites and ZDRs track along the pattern of known and predicted genes in human chromosome 22 [40] (Fig. 6). In addition, unlike the *E. coli* genome, both accumulate at the TSS of genes (with the peak for ZDRs just upstream and that for NFI sites just down stream of the TSS). Finally, NFI sites, along with GC content were suppressed at the 3′-ends of genes (ZDRs were flat in this region, but this is because there is near zero ZDRs throughout the entire length of human genes, except at the TSS). The similarity in thermogenomic behavior of ZDRs with a known eukaryotic promoter (the NFI sites) further supports the role of ZDRs in transcription regulation of human genes, and the architecture with ZDRs peaking just upstream of NFI sites around the TSS suggesting that the coactivation model for CSF-1 as proposed by Liu et al. [19] may be fairly general for human genes.

4. Phylogenomic analyses

A genomic analysis allows us to identify within a genome where a particular structure or binding site has a high probability of occurring and, therefore, the potential function for such sites within an organism. The functional significance of such elements would be bolstered if they consistently show similar patterns of occurrence across similar organisms, or differ in organisms that are not very similar. For example, we would expect the distribution of NFI sites relative to genes to be consistent between eukaryotic species, but differ from that of prokaryotes. This requires a phylogenomic analysis, comparing the occurrence of such markers of genomes across species, phyla, or even kingdoms. The important factor here, as with genomic analyses, is to insure accurate annotations of the functional features in the genomes (distinguishing genes from predicted genes, from pseudogenes, etc.). In the case of a phylogenomic analysis, it is also critical that the annotations are consistent across organisms. In our studies, we used only the annotations available in the ENSEMBL database for eukaryotic genomes (www.ensembl.org). This means, however, that some organisms will not be represented, as their genomes are annotated separately.

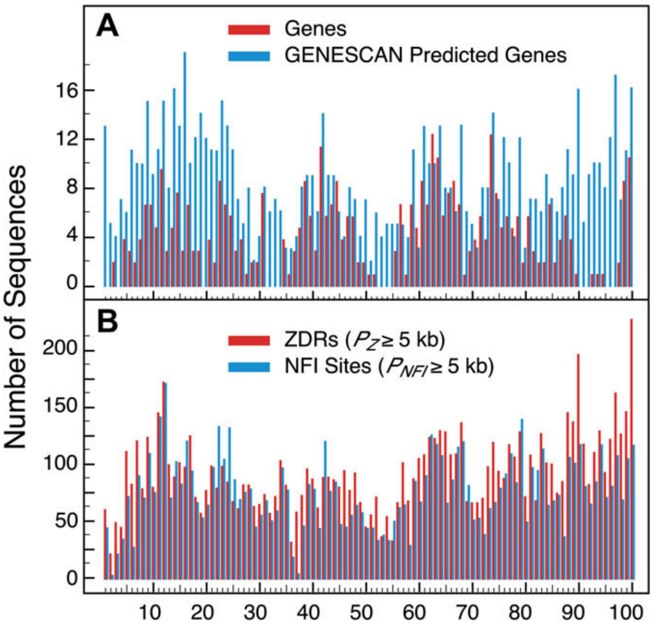

Fig. 6. Distribution of genes, ZDRs, and NFI sites along human chromosome 22 [40]. (A) The distribution pattern of genes and predicted GENSCAN predicted genes are shown as red and blue bars, respectively. (B) The distributions of ZDRs ($P_Z \geqslant 700$ bp) and NFI sites ($P_{NFI} \geqslant 500$ bp) are shown as red and blue bars, respectively.

An obvious comparison across organisms is for the occurrence of NFI sites in eukaryotic *versus* prokaryotic genomes—as a eukaryotic transcriptional regulator, one would expect differences in the pattern of occurrence for such sites in the two different classes of organisms. Indeed, such an analysis showed that, although nearly every gene, pro- or eukaryotic, contained at least one NFI site, it was the number of NFIs within genes and near the TSS that distinguishes these two major classes of organisms. In prokaryotes, a very large percentage of NFIs were in genes and near the TSS of genes, while, in eukaryotes, these numbers were significantly lower, and decreased rapidly with increased complexity of the organism. This would initially appear to be contrary to what one would expect; however, what this demonstrates is the concept of discrimination in the localization of such sequences. Looking back at the human CSF-1 gene (Fig. 5), for example, the potentially functional NFI sites are spread out across the gene, but only one occurred at within 1 kb of the TSS. Thus, the system requires only this one additional signal (that of a strong ZDR) to discriminate between the truly functional and other high affinity NFI sites. ZDRs show a similar pattern, with the number of ZDRs in genes and near the TSS becoming increasingly discriminate with increasing complexity of the organism. Again, the near identical patterns of occurrence between NFIs and ZDRs suggest a functional role of potential Z-DNA sequences in transcriptional regulation.

The other questions that can be readily addressed using a phylogenomic analysis are those concerning the evolution of sequences, for example, how such elements emerged relative to one another. In our particular studies, we were interested in determining whether coactivation by NFI and ZDRs resulted from parallel or convergent evolution. An analysis of the occurrence of NFIs relative to the TSS of genes in organisms from eubacteria to archaebacteria to simple eukaryotes, to complex eukaryotes, to animals indicated that such sites emerged as a functional transcriptional regulator from the general accretion of GC-rich sequences in eukaryote genes, and the increasing complexity of the transcriptome, which resulted in the migration of the TSS downstream of the AT-rich promoters and into the gene itself [47] (Fig. 7). ZDRs, however, emerged as two separate classes: the first is the CG-rich ZDRs (exemplified by the standard alternating CpG type Z-DNA sequences), which followed the pattern of emergence as NFIs and other GC-rich elements; the second are AT-rich type ZDRs (such as the CpA/TpG type promoters), which emerged upstream of the TSS, but migrated with the TSS until they converged with other GC-rich elements, including NFIs. Thus, if we consider that most ZDRs that have been identified as eukaryotic promoters are of the CpA/TpG-type, we can conclude that the mechanism of coactivation seen in the human CSF-1 gene results from convergence of two different markers.

One issue that must be addressed in searching for Z-DNA and NFI sites is that both are GC-rich elements and, therefore, we need to be concerned with the effect of nucleotide composition on the results of the analyses. Thus, the measures for probabilities (P_Z and P_{NFI}) are relative to random sequences with the average GC-content of eukaryotic systems (the weighted average of organisms studied = 42.3%).

5. Perspectives

Two applications of thermogenomics are described in detail as they apply to genomic and phylogenomic analyses of DNA se-

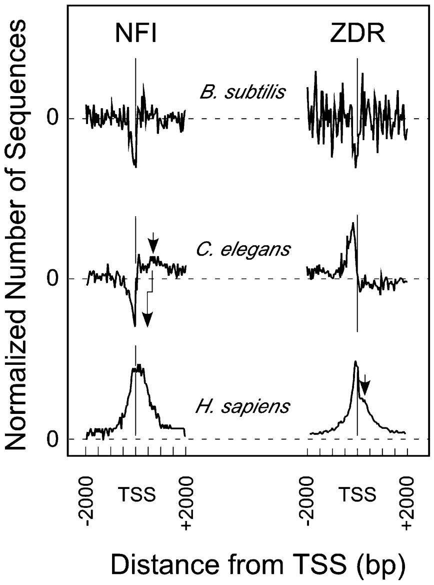

Fig. 7. Comparison of the distribution patterns for ZDRs and NFIs around the transcription start-sites (TSS) of *B. subtilis*, *C. elegans*, and human genes [47].

quences with the potential to adopt left-handed Z-DNA or bind to the eukaryotic nuclear factor I regulatory element. It should be noted that such thermogenomic analyses identify sequences with high propensities to function in these manners, not that they necessarily do function in this way for any particular gene or organism. The caution, therefore, is that the results of such analyses, at least at the gene and organism level, are meant primarily to lead the design of further confirmatory studies. That being said, the power of the thermogenomic approach is that there is a strong correlation between the predicted propensities and the actual abilities of the sequences to function in their respective manners. More importantly, the analyses do not rely solely on the search for “consensus” sequences, but can accommodate a large number of variations from stringently defined sequence motifs. As such, the approach is applicable to any type of structural transformation or protein binding motif, assuming that the thermodynamic parameters are available. For example, it may not be necessary to perform the systematic studies of Roulet et al. [46] in order to establish the binding scores for a protein and its cognate DNA. One can easily imagine that, through a SELEX approach, one can determine the statistical probability of finding a particular binding motif and its variations from a set of random sequences and, from that, derive the effect of such variations on the affinity constants, which would form the basis for the development of a thermogenomic algorithm for that protein.

The application of thermogenomics approaches can obviously be extended to studying other types of DNA structures and their functions. ZHUNT, for example, is currently being extended to include predicting the effects of cytosine methylation on the propensity for Z-DNA formation in genomes. There are also very detailed structural as well as thermodynamic data available for the extrusion of cruciforms, another conformation that can be studied in terms of potential function through a thermogenomic approach. Sequence motifs and the effect of methylation have previously been identified for structures that promote spontaneous cytosine deamination [49] and the stability of Holliday junctions [50,51], which can potentially be applied to predicting hot spots for transition mutations and recombination events, respectively. Finally, the statistical occurrence of DNA sequences that are associated with RNA structures and their thermodynamic properties (hairpins [52,53] and G-quartets [54,55]) may lead to better understanding of how genes are regulated beyond the transcriptional level, including mRNA splicing and regulation of translation.

In summary, a thermogenomic search is much more complex than searches for simple sequence motif. However, through the two examples described here, the additional effort has the potential to provide accurate predictions for not only where such sequences may occur, but also quantitative measures that can be used to accurately compare the propensities for structure formation or protein binding that can lead to better address questions concerning the functional or evolutionary significance of a DNA sequence.

References

[1] J.D. Watson, F.H. Crick, Nature 171 (1953) 737–738.
[2] J.F. Riou, L. Guittat, P. Mailliet, A. Laoui, E. Renou, O. Petitgenet, F. Megnin-Chanet, C. Helene, J.L. Mergny, Proc. Natl. Acad. Sci. USA 99 (2002) 2672–2677.
[3] C.W. Greider, Bioessays 12 (2005) 363–369.
[4] L. Oganesian, T.M. Bryan, Bioessays 29 (2007) 155–165.
[5] M. Mills, L. Lacroix, P.B. Arimondo, J.L. Leroy, J.C. Francois, H. Klump, J.L. Mergny, Curr. Med. Chem. Anticancer Agents 2 (2002) 627–644.
[6] A. Ghosh, M. Bansal, Acta Crystallogr. D Biol. Crystallogr. 59 (2003) 620–626.
[7] S. Neidle, Oxford Handbook of Nucleic Acid Structure, Oxford Press, Oxford, 1999.
[8] S.M. Mirkin, Front Biosci. 13 (2008) 1064–1071.
[9] A.H. Wang, G.J. Quigley, F.J. Kolpak, J.L. Crawford, J.H. van Boom, G. van der Marel, A. Rich, Nature 282 (1979) 680–686.
[10] H.R. Drew, R.M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura, R.E. Dickerson, Proc. Natl. Acad. Sci. USA 78 (1981) 2179–2183.
[11] J. Marx, Science 230 (1985) 794–796.
[12] A. Rich, S. Zhang, Nat. Rev. Genet. 4 (2003) 566–572.
[13] E. Pennisi, Science 312 (2006) 1467–1468.
[14] M. Morange, J. Biosci. 32 (2007) 657–661.
[15] A. Herbert, A. Rich, Proc. Natl. Acad. Sci. USA 98 (2001) 12132–12137.
[16] J.A. Kwon, A. Rich, Proc. Natl. Acad. Sci. USA 102 (2005) 12759–12764.
[17] J. Nickol, M. Behe, G. Felsenfeld, Proc. Natl. Acad. Sci. USA 79 (1982) 1771–1775.
[18] J. Ausio, G. Zhou, K. van Holde, Biochemistry 26 (1987) 5595–5599.
[19] R. Liu, H. Liu, X. Chen, M. Kirby, P.O. Brown, K. Zhao, Cell 106 (2001) 309–318.
[20] P.S. Ho, M.J. Ellison, G.J. Quigley, A. Rich, EMBO J. 5 (1986) 2737–2744.
[21] G.P. Schroth, P.J. Chou, P.S. Ho, J. Biol. Chem. 267 (1992) 11846–11855.
[22] T.L. Bailey, Methods Mol. Biol. 452 (2008) 231–251.
[23] G.P. Schroth, P.S. Ho, Nucleic Acids Res. 23 (1995) 1977–1983.
[24] P.R. Hoyne, L.M. Edwards, A. Viari, L.J. Maher 3rd, J. Mol. Biol. 302 (2000) 797–809.
[25] F. Lillo, S. Basile, R.N. Mantegna, Bioinformatics (Oxford, England) 18 (2002) 971–979.
[26] M.P. Kladde, Y. Kohwi, T. Kohwi-Shigematsu, J. Gorski, Proc. Natl. Acad. Sci. USA 91 (1994) 1898–1902.
[27] R. Cox, S.M. Mirkin, Proc. Natl. Acad. Sci. USA 94 (1997) 5237–5242.
[28] M.A. Fuertes, V. Cepeda, C. Alonso, J.M. Perez, Chem. Rev. 106 (2006) 2045–2064.
[29] S.C. Ha, K. Lowenhaupt, A. Rich, Y.G. Kim, K.K. Kim, Nature 437 (2005) 1183–1186.
[30] F. Azorin, R. Hahn, A. Rich, Proc. Natl. Acad. Sci. USA 81 (1984) 5714–5718.
[31] B.H. Johnston, A. Rich, Cell 42 (1985) 713–724.
[32] A. Nordheim, L.J. Peck, E.M. Lafer, B.D. Stollar, J.C. Wang, A. Rich, Cold Spring Harb. Symp. Quant. Biol. 47 (Pt. 1) (1983) 93–100.
[33] J.C. Wang, L.J. Peck, K. Becherer, Cold Spring Harb. Symp. Quant. Biol. 47 (Pt. 1) (1983) 85–91.
[34] L.J. Peck, J.C. Wang, Proc. Natl. Acad. Sci. USA 80 (1983) 6206–6210.
[35] U. Egner, J. Kratzschmar, B. Kreft, H.D. Pohlenz, M. Schneider, Chembiochem 6 (2005) 468–479.
[36] M.D. Frank-Kamenetskii, A.V. Vologodskii, Nature 307 (1984) 481–482.
[37] A.V. Vologodskii, M.D. Frank-Kamenetskii, J. Biomol. Struct. Dyn. 1 (1984) 1325–1333.
[38] M.J. Ellison, J. Feigon, R.J. Kelleher 3rd, A.H. Wang, J.F. Habener, A. Rich, Biochemistry 25 (1986) 3648–3655.
[39] M.J. Ellison, R.J. Kelleher 3rd, A.H. Wang, J.F. Habener, A. Rich, Proc. Natl. Acad. Sci. USA 82 (1985) 8320–8324.
[40] P.C. Champ, S. Maurice, J.M. Vargason, T. Camp, P.S. Ho, Nucleic Acids Res. 32 (2004) 6501–6510.
[41] P.S. Ho, Proc. Natl. Acad. Sci. USA 91 (1994) 9549–9553.
[42] K. Nagata, R.A. Guggenheimer, T. Enomoto, J.H. Lichy, J. Hurwitz, Proc. Natl. Acad. Sci. USA 79 (1982) 6438–6442.
[43] R.M. Gronostajski, Nucleic Acids Res. 14 (1986) 9117–9132.
[44] K.A. Jones, J.T. Kadonaga, P.J. Rosenfeld, T.J. Kelly, R. Tjian, Cell 48 (1987) 79–89.
[45] S. Osada, S. Daimon, T. Nishihara, M. Imagawa, FEBS Lett. 390 (1996) 44–46.
[46] E. Roulet, P. Bucher, R. Schneider, E. Wingender, Y. Dusserre, T. Werner, N. Mermod, J. Mol. Biol. 297 (2000) 833–848.

[47] P. Khuu, M. Sandor, J. DeYoung, P.S. Ho, Proc. Natl. Acad. Sci. USA 104 (2007) 16528–16533.

[48] C. Burge, S. Karlin, J. Mol. Biol. 268 (1997) 78–94.

[49] J.M. Vargason, B.F. Eichman, P.S. Ho, Nat. Struct. Biol. 7 (2000) 758–761.

[50] B.F. Eichman, J.M. Vargason, B.H.M. Mooers, P.S. Ho, Proc. Natl. Acad. Sci. USA 97 (2000) 3971–3976.

[51] F.A. Hays, A. Teegarden, Z.J. Jones, M. Harms, D. Raup, J. Watson, E. Cavaliere, P.S. Ho, Proc. Natl. Acad. Sci. USA 102 (2005) 7157–7162.

[52] I. Toulokhonov, I. Artsimovitch, R. Landick, Science 292 (2001) 730–733.

[53] I. Toulokhonov, R. Landick, Mol. Cell. 12 (2003) 1125–1136.

[54] M.C. Didiot, Z. Tian, C. Schaeffer, M. Subramanian, J.L. Mandel, H. Moine, Nucleic Acids Res. 36 (2008) 4902–4912.

[55] S. Bonnal, C. Schaeffer, L. Creancier, S. Clamens, H. Moine, A.C. Prats, S. Vagner, J. Biol. Chem. 278 (2003) 39330–39336.

[56] S.E. Warne, P.L. deHaseth, Biochemistry 32 (1993) 6134–6140.

[57] H.J. Lipps, W. Gruissem, D.M. Prescott, Proc. Natl. Acad. Sci. USA 79 (1982) 2495–2499.

[58] W.I. Sundquist, A. Klug, Nature 342 (1989) 825–829.

[59] D. Sen, W. Gilbert, Nature 344 (1990) 410–414.

[60] G. Wang, K.M. Vasquez, Proc. Natl. Acad. Sci. USA 101 (2004) 13448–13453.

[61] D.M.J. Lilley, Quart. Rev. Biochem. 33 (2000) 109–159.

[62] S. Ahmed, A. Kintanar, E. Henderson, Nat. Struct. Biol. 1 (1994) 83–88.

[63] J.L. Leroy, M. Gueron, J.L. Mergny, C. Helene, Nucleic Acids Res. 22 (1994) 1600–1606.

[64] A.T. Phan, M. Gueron, J.L. Leroy, J. Mol. Biol. 299 (2000) 123–144.

[65] G. Wang, K.M. Vasquez, Front Biosci. 12 (2007) 4424–4438.

[66] G.N. Parkinson, M.P. Lee, S. Neidle, Nature 417 (2002) 876–880.