2004

Distributions of Z-DNA and Nuclear Factor I in Human Chromosome 22: A Model for Coupled Transcriptional Regulation

P. Christoph Champ

Sandor Maurice

Jeff Vargason
George Fox University, jvargason@georgefox.edu

Tracy Camp

P. Shing Ho
hops@onid.orst.edu

Recommended Citation
Previously published in Nucleic Acids Research, 2004, 32, pp. 6501-6510 http://nar.oxfordjournals.org/content/32/22/6501

Published online December 14, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 22 ***6501–6510***
doi:10.1093/nar/gkh988

Distributions of Z-DNA and nuclear factor I in human chromosome 22: a model for coupled transcriptional regulation

P. Christoph Champ, Sandor Maurice, Jeffrey M. Vargason, Tracy Camp and P. Shing Ho*

Department of Biochemistry and Biophysics, ALS 2011, Oregon State University, Corvallis, OR 97331, USA

Received as resubmission July 12, 2004; Revised and Accepted November 22, 2004

ABSTRACT

An analysis of the human chromosome 22 genomic sequence shows that both Z-DNA forming regions (ZDRs) and promoter sites for nuclear factor-I (NFI) are correlated with the locations of known and predicted genes across the chromosome and accumulate around the transcriptional start sites of the known genes. Thus, the occurrence of Z-DNA across human genomic sequences mirrors that of a known eukaryotic transcription factor. In addition, 43 of the 383 fully annotated chromosomal genes have ZDRs within 2 nucleosomes upstream of strong NFIs. This suggests a distinct class of human genes that may potentially be transcriptionally regulated by a mechanism that couples Z-DNA with NFI activation, similar to the mechanism previously elucidated for the human colony stimulation factor-I promoter [Liu *et al.* (2001) *Cell*, 106, 309–318]. The results from this study will facilitate the design of experimental studies to test the generality of this mechanism for other genes in the cell.

INTRODUCTION

The biological relevance of Z-DNA has been controversial since its discovery in 1979 (1), with many of the early studies raising as many questions as they answered (2). In recent years, more rigorous studies have started to reveal specific roles for Z-DNA in processes such as RNA editing (3,4). One particularly interesting study of genes regulated by BAF, the mammalian analogue of the yeast SWI/SNF switching complex, has resurrected the possible role of Z-DNA as a transcriptional regulator when coupled with the nuclear factor-I (NFI) or CAAT-box transcription factor (5). Here, we use a computer-assisted approach to identify sites in human chromosome 22 that have the potential to form Z-DNA and that are recognized as NFI promoter sites. This allows us to determine whether Z-DNA coupled transcriptional-regulation may be more general to a broader class of human genes.

Left-handed Z-DNA has been attributed to a variety of biological functions in the cell [reviewed in (6)]. For many years, claims of a biological role for Z-DNA were greeted with great skepticism (2). Recently, however, the DNA binding domains of double-stranded RNA adenosine deaminases have been shown to bind specifically to Z-DNA (3,4,7,8). The participation of Z-DNA in RNA processing is consistent with the observation that Z-DNA can be induced at the 5′ ends of genes by transcription (9,10), and that potential Z-DNA forming sequences accumulate at and near the transcriptional start sites (TSS) of human genes (11). However, the large number of genes predicted to contain potential Z-forming sequences (11) is inconsistent with such a limited role.

Liu *et al.* (5) have recently revived the possibility that Z-DNA may be an important component in the regulation of certain mammalian genes. In their study, a DNA array was used to screen a library of genes that are regulated by the mammalian BAF complex. The screen identified 80 mRNA sequences whose levels are elevated and 2 that are suppressed by the BAF complex. A detailed biochemical analysis of the colony stimulating factor-I (CSF-1) gene, associated with one of the induced mRNAs, revealed a classic Z-forming alternating CA/TG-repeat sequence immediately upstream of the NFI consensus site. Replacing this CA/TG-repeat with a random sequence was seen to suppress the activity of the CSF-1 promoter. Furthermore, this CA/TG-repeat was shown to be left-handed when the gene is actively being transcribed. A model was proposed (Figure 1) in which Z-DNA maintains the CSF-1 promoter in an activated state (when coupled with the activity of a transcriptional activator) by maintaining an open chromatin structure—it is well known that Z-DNA does not bind to nucleosomes (12–15).

Here, we address the question of whether this mechanism of coupling NFI activation with transcription-induced Z-DNA may be unique to the CSF-1 gene or characteristic of a larger class of similarly regulated genes. In the current study, sequences that have strong thermodynamic potential to form Z-DNA and to bind NFI were identified in human chromosome 22 (16,17), a small genomic sequence with a relative high density of genes. The results indicate that a well-defined group of human genes may be regulated

*To whom correspondence should be addressed. Tel: +541 737 2769; Fax: +541 737 0481; Email: hops@onid.orst.edu
Present address:
Jeffrey M. Vargason, National Institute of Environmental Health Sciences, MD F3-05, PO Box 12233, Research Triangle Park, NC 27709, USA

by coupling Z-DNA with a eukaryotic transcription factor. Similar analyses of the *Escherichia coli* (*E.coli*) genome (18) shows that this mechanism is not present in prokaryotes.

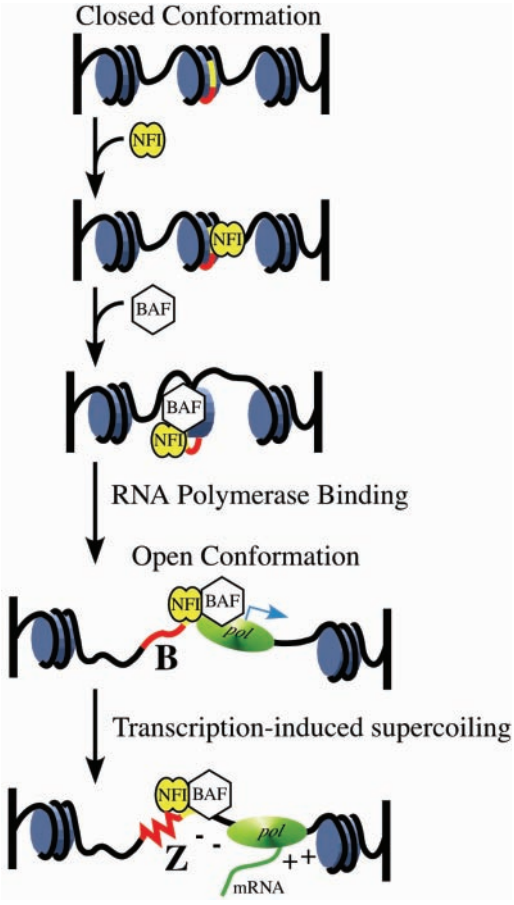

Figure 1. Proposed mechanism for Z-DNA coupled transcriptional regulation in BAF regulated genes (5). In this model, a nuclear factor I (NFI) consensus sequence (yellow) in an inactive compact chromatin structure is first bound by the transcription factor. The closed conformation of the chromatin is relaxed by binding of the BAF complex and is further relaxed to the fully active open conformation upon recruitment of RNA polymerase. Transcription of the gene generates positive supercoils in front and negative supercoils behind the polymerase in the topologically constrained chromatin fiber (22). The negative supercoils induce a transition to left-handed Z-DNA in an upstream region (red) (9), which maintains the chromatin in its activated open conformation.

METHODS

Sequence data

The sequence of chromosome 22 q-arm, release 3.1b (March 5, 2002), was retrieved as a single FASTA sequence from the Sanger lab site (http://www.sanger.ac.uk/) (Table 1). The definitions of genes, pseudogenes and predicted genes (from GENSCAN), along with start and end of each human gene, were defined according to annotations in the Sanger site (complete as of March 25, 2003), with the direction and start site of transcription determined by alignment of a gene sequence with its associated mRNA or cDNA sequence. The sequence of the *E.coli* strain K-12 was used as the representative prokaryotic genome, and retrieved as a single file along with annotations from the NCBI website (ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coli_K12/).

Defining Z-DNA regions (ZDRs)

We had previously mapped Z-DNA in genes and genomes at the base pair level using the program ZHUNT (11,19). Such a detailed map in very large genomic DNAs, however, is burdensome; therefore, in the current study, contiguous nucleotides with strong potentials to form Z-DNA are grouped into Z-DNA regions, or ZDRs. A ZDR is defined as a set of contiguous base pairs with Z-DNA propensities (P_Z) that are greater than or equal to a predefined lower limit (P_{Zmin}). The parameter P_Z is defined here, called the 'Z-score' in earlier versions of ZHUNT (11,19), as the number of random base pairs that must be searched to find a sequence that is as good or better at forming Z-DNA as the sequence being analyzed. P_Z, therefore, is a statistical test for the uniqueness of a sequence. Here, we have recalculated this parameter to reflect the proper distribution of the probabilities for any sequence within the context of 33 million random nucleotides; therefore P_Z is now more accurately defined, and is distinguished from the true statistical '*z*-score', which is calculated differently. In searching through long genomic sequences, a new ZDR is started with a base pair in which $P_Z \geqslant P_{Zmin}$ and is terminated with any base pair in which $P_Z < P_{Zmin}$. The minimum size of a ZDR is 12 bp, or one full turn of Z-DNA.

We set the value for $P_{Zmin} = 0.5$ kb. This corresponds to a contiguous stretch of alternating CA/TG 12 bp in length, which has been shown experimentally to adopt Z-DNA under reasonable levels of negative superhelical densities (20), and conforms to our definition of a minimum Z-forming sequence in previous studies (11,19). The archetypical Z-DNA

Table 1. Occurrence of ZDRs in human chromosome 22 (16,17) and the *E.coli* genome (18)

Sequence type	Number	Number of bp	Number of sequences w/ZDRs	% Sequences w/ZDRs	Total number of ZDRs	Expected # ZDRs[a]
Chromosome 22		33 821 688			7580	(48 317)
Genes	383	15 671 983	370	96.6%	4009[b]	3684
Pseudogenes	234	837 815	100	42.7%	313[b]	293
Predicted genes	790	22 887 969	687	87.0%	5969[b]	5484
E.coli genome		4 639 262			3500	(6627)
Genes	4279	4 098 295	3597	82.7%	3497	3092

[a]Number of ZDRs predicted for a sequence (in parentheses) was calculated by taking the total number of base pair in that sequence and dividing by P_{min} (500 bp). The predictions for how the observed ZDRs are distributed between genes, pseudogenes and predicted genes are based on the 7580 actually observed in chromosome 22 multiplied by the fraction of the total sequence represented by that sequence type.
[b]ZDRs that appear in two different overlapping gene, pseudogene and/or predicted gene sequences are counted for both sequences.

sequence of 24 bp, d(CG)$_{12}$ has the highest possible P_Z value (4.1×10^8); there are, however, no alternating d(CG) of such length in this chromosomal sequence. A simple sequence search for this motif therefore could not constitute a comprehensive search for the occurrence of the left-handed conformation.

Defining potential binding sites for the NFI transcription factor

NFI recognizes the consensus inverted repeat sequence TTGGCN$_5$GCCAA, where N$_5$ is a spacer of any 5 nt. Roulet *et al*. (21) have systematically determined the effect on NFI binding of modifying any nucleotide and the spacer length within this consensus sequence. The resulting set of binding scores can thus be used to predict the relative binding affinities for NFI (Figure 2a). The actual variations in these NFI binding scores, however, are small. Even the known NFI promoter sequence in, for example, the CSF-1 gene (5) does not distinguish itself above the noise. In order to define a propensity for binding that is analogous and thus comparable to the propensities for Z-DNA formation (P_Z), we have applied a statistical test for the occurrence of all possible variations of the NFI consensus sequence.

We have defined a propensity for a sequence to bind NFI (P_{NFI}) in a manner analogous to P_Z, where P_{NFI} is the number of random sequences that must be searched to find a site that has the same or higher affinity for the transcription factor. For this parameter, we first calculated the NFI binding scores for all possible variations of the consensus full and half sites. This allowed us to define explicitly the frequency and thus uniqueness of every possible NFI binding score as defined by Roulet *et al*. (21). In the case of the CSF-1 gene (Figure 2b), the known NFI promoter stands-out with a $P_{NFI} \approx 1.2 \times 10^5$ bp (equivalent to a raw binding score of 98), which reflects the uniqueness of this sequence. We should note that P_{NFI} values mirror the affinity constant K_A, except for the magnitude of the values (P_{NFI} values are ~2000 times smaller than the comparable K_A). Thus, P_{NFI} calculated in this manner provides a conceptual link between the affinity and the uniqueness of any particular nuclear transcription factor recognition site.

With the P_{NFI} values, we defined criteria to help distinguish functional from nonfunctional NFI sites. It is clear that binding affinity is not the sole determinant of activity, since a sequence with a binding score of 80 can show greater activation than one with a score of 84 (21). If we assume that the function of NFI as a eukaryotic transcription factor should distinguish itself in human sequences, then we should see a distinction between the occurrence of NFI sites between chromosome 22 and a similar random sequence. For this analysis, we generated a random sequence with the same length and G + C content as that of human chromosome 22, and mapped the location of genes from the chromosome onto this random sequence. We then analyzed the human and random sequences for NFI sites within the defined gene regions for various values of P_{NFImin} and with various lengths of sequence added to both sides of the TSS of each gene. The differences between the number of NFIs located at the gene positions of the random chromosome sequence and those of identified genes in the actual chromosomal sequence are plotted for each P_{NFI} value (Figure 3a). From this analysis, we see that the human sequences become distinct from random (difference > 0) for P_{NFI} values ⩾0.5 kb

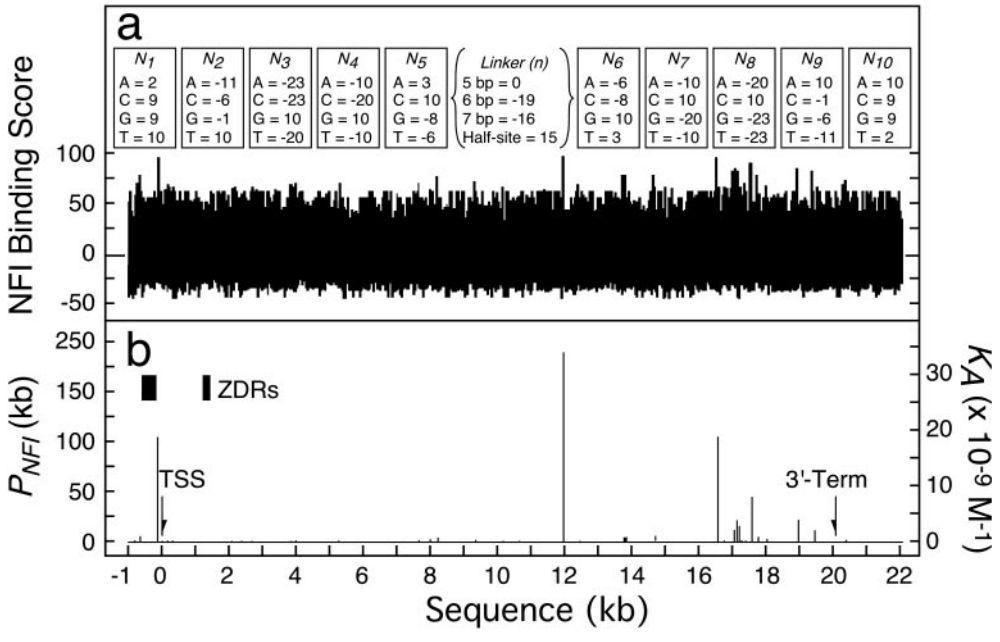

Figure 2. Prediction of NFI and ZDRs in the human CSF-1 gene. (**a**) Binding scores as defined by Roulet *et al*. (21). The binding scores are calculated according to values assigned to each nucleotide position and linker length (*n*), with the maximum score = 100 for the NFI consensus sequence. The best score was calculated starting with each nucleotide of the CSF-1 gene, including 1 kb of sequence upstream of the transcriptional start site (TSS) and 2 kb downstream of the 3′-terminus. (**b**) Comparison of the statistical uniqueness (P_{NFI}) and association constant (K_A) for NFI sites, and locations of ZDRs. Binding scores for each nucleotide are converted to P_{NFI} values (in kb) from A, as described in Methods, and compared to the analogous K_A values. The locations of ZDRs along the gene, as predicted by ZHUNT, are shown as bars.

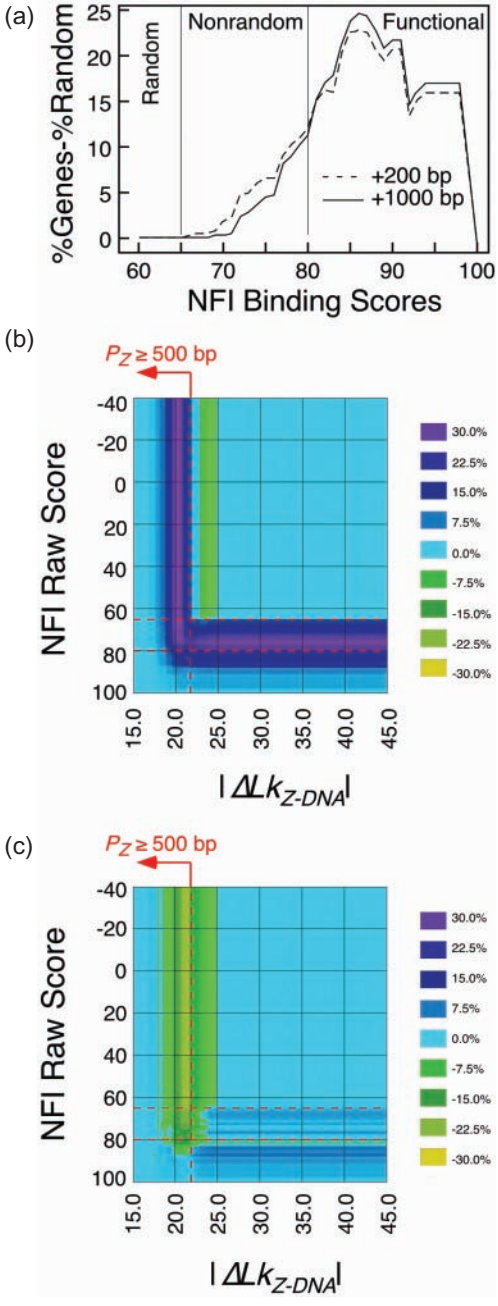

Figure 3. Classification of ZDRs and NFI sites. (**a**) The percentage of gene sequences with NFI sites located in chromosome 22 and in a random sequence with the same length and C + G content as this chromosome (with gene locations mapped from the chromosomal sequence) are subtracted and compared to the raw binding scores of Roulet *et al*. (21) (*x*-axis). The chromosomal genes deviate from the random sequence at scores ≥65 (or at $P_{NFI} \geq 0.5$ kb), and show the greatest deviation from random at NFI binding scores of 87. Binding scores at the inflection point for the curve (≥80, or equivalent to $P_{NFI} \geq 5$ kb) define a class of strong and potentially functional NFI sites. Sequences with binding scores between 65 and 80 (0.5 kb ≤ P_{NFI} ≤ 5 kb) define the nonrandom but weak binding NFI sites. (**b**) The percentage difference in the number of ZDRs and NFI sites located in human chromosome 22 relative to random DNA with overall C + G content identical to that of the chromosomal sequence. The *x*-axis shows an increasing number of supercoils required to induce Z-DNA ($|\Delta Lk_{Z\text{-}DNA}|$, associated with increasingly poor Z-DNA forming sequences), while the *y*-axis are the NFI scores calculated using the values reported by Roulet *et al*. (21). (**c**) The percentage difference in the number of ZDRs and NFI sites located in human chromosome 22 and a random sequence with the identical C + G distribution across the sequence as the chromosome.

(or a raw binding score of 65), with the largest deviation from random occurring at $P_{NFI} \approx 20$ kb (raw score of 86). We have defined $P_{NFI} \approx 5$ kb (raw score of 80), midway between random and highly nonrandom, to distinguish a class of strong and potentially functional NFI sites. In comparison, there was no difference seen for NFI sites located in *E.coli* genes as compared to the associated random sequences, indicating that, as expected, eukaryotic transcription factor consensus sequences are located randomly in a prokaryotic genome. We therefore have classified NFI sites as random P_{NFI} values ≤0.5 kb, nonrandom and weak P_{NFI} values between 0.5 and 5 kb, and strong P_{NFI} values ≥5 kb, and will use these classifications for the remainder of this paper. These definitions are consistent with the functional activities reported for NFI promoter sites (21). Sequences that are categorized as nonrandom and weak include those that show ∼10% of the activity seen with a consensus NFI sequence in a luciferase assay, but significantly above background, while those categorized as strong and potentially functional exhibit at least 25% of the activity compared to the consensus sequence. Finally, all but one of the 37 annotated genes identified as BAF regulated (5) have at least one NFI site identified with $P_{NFI} \geq 0.5$ kb.

To test for the accuracy of the predictions for promoter binding, we located 28 genes where there is good experimental evidence for protein binding at an NFI promoter or its variant (www.gene-regulation.com). Of these, 26 (92.9%) were found by our program to have sites with raw binding scores ≥65 ($P_{NFI} \geq 0.5$ kb, our definition for a nonrandom sequence) within 1 kb of the identified TSS, and 10 of these had scores ≥80 (defined here as being functional) (Supplementary Table 1). In contrast, it is almost impossible to test for false positives for this class of transcriptional factors. The most definitive test would be to apply the program towards genes that are known to not bind NFI protein; however, no such dataset exists to our knowledge. For example, many genes with active *cis*-regulatory elements will also have an NFI/CTF binding site, and may be either positively or negatively regulated by these sites. Thus, we cannot simply assemble a dataset of genes that are known to be regulated by a promoter other than NFI, since there is no experimental evidence that such genes do not also include functional NFI binding sites. Thus, although we are very confident that the cutoffs defined here are sufficient to include nearly all genes known to bind the NFI protein, we cannot make any accurate estimates for the number of genes predicted to have an NFI binding site that does not bind protein in the cell.

RESULTS

In this study, we address the question of whether there is any relationship between NFI activation and the occurrence of Z-DNA in human genes. The sequence of human chromosome 22 was analyzed using the statistical thermodynamic treatment for Z-DNA implemented in the ZHUNT program (11,19), which was previously shown to accurately predict sequences with strong potential to adopt the left-handed conformation, and an analogous statistical method based on the affinity constants (21) used to search for potential NFI binding sites. Applying thermodynamic strategies to identify potentially

functional sequences has both advantages and disadvantages. The advantage is that there is no need for training sets. If we accept, for example, that the affinity of a nuclear factor for a particular site plays a major role in transcriptional regulation, then a search for all sites with affinity constants that allow for recognition and binding should yield a list of all DNA sequences that could serve as potential promoter sites. The disadvantage, of course, is that we need to define what are considered to be strong or relevant affinity constants, and we do not consider other factors (such as interactions with accessory proteins) in the thermodynamic model that may be important for defining a functional promoter. In the current study, we use the thermodynamic algorithm in ZHUNT to identify potential Z-DNA forming sequences (ZDRs). The algorithm in ZHUNT had previously been shown to accurately define not only the location at the nucleotide level, but also the relative propensity of sequences in genomes to adopt the left-handed Z-conformation (19). The actual formation of Z-DNA in these sequences in the cell requires introduction of negative supercoiling (e.g. during chromatin remodeling) and, therefore, we should recognize that the potential ZDRs will become left-handed only under specific circumstances, which are not entirely well understood and which we do not attempt to predict here. Similarly, for the NFI sites, promoters are defined by many criteria other than affinity of a protein for a sequence. Thus, although our general definitions follow the functional luciferase assay results as originally reported by Roulet *et al*. (21), we should again qualify the search results by indicating that these are all potential promoter sites.

We can ask how the cut-off values of the two sequence types (ZDRs and NFI sites) correspond to what one would expect around the TSS of genes. Figure 3b compares the occurrence of ZDRs and NFI sites at various values of superhelical density (ΔLk) required to induce a ZDR to become left-handed (as predicted from ZHUNT) and the raw binding scores for NFI recognition of sites for sequences 1000 bp upstream and downstream of the TSS for the annotated genes of chromosome 22 relative to a random chromosome. In this analysis, the random chromosome is generated with the same length and overall average C + G content as the human sequence, with the locations of the transcriptional start and end sites, and direction of transcription mapped onto this random sequence. This comparison shows that for the $P_{\mathrm{Zmin}} \geqslant 500$ bp ($\Delta Lk \geqslant -22$ supercoils), the number of ZDRs located is significantly higher (up to 30% greater) than expected in the random sequence. Similarly, for NFI scores $\geqslant$60, the number of potential promoter binding sites is greater than in the random sequence. At very high values of P_{Zmin} and NFI scores, what is observed in the human genes approaches what is expected for the random sequence, probably because of the very small numbers of sequences one would expect.

If we take the C + G content in and around the TSS into account (generating random sequences with the same distributions of C + G content at each position either upstream or downstream of the TSS), the numbers of ZDRs is again nonrandom, but becomes lower than expected. For potential NFI binding sites, there are apparently two sets of nonrandom groups of sequences, one between scores of 60 and 80 and the other with NFI scores >80. Thus the nonrandom behaviour of the ZDRs for $P_{\mathrm{Zmin}} \geqslant 500$ bp is consistent with what we had expected from the types of sequences that can reasonably form Z-DNA. In addition, the two classes of nonrandomly behaving NFI type sequences with binding scores from 65 to 80 and $\geqslant$80, which we define as nonrandom and strong potential NFI sites, respectively, are consistent with what was reported in terms of the effect of such sequences on gene activation according to a luciferase assay.

Although ZHUNT had previously been applied to a limited set of human genes, the current analysis allows us to identify ZDRs in various functional sequence classes in a complete chromosome. The corresponding analysis for potential NFI sites reveals the proportion of genes in chromosome 22 that are potentially regulated by this nuclear transcription factor. Finally, comparing the results of the two analyses allow us to answer the question of whether Z-DNA is associated with a particular class of regulated genes and, thus, plays a general function in the regulation of transcription for a distinct group of human genes. With the qualifications stated above, the list of gene sequences that result from this analysis should be interpreted as being a starting point to help the experimentalist design studies to determine whether a particular gene actually utilizes such a coupled mechanism of transcriptional regulation.

Occurrence of NFIs and ZDRs in human chromosome 22 and the *E.coli* genome

There are a very large number of NFI sites in the nonrandom and functional classes identified in the human chromosome 22 sequence (Table 2), slightly higher than expected from random. In addition, nearly every sequence identified and predicted as a potential gene in human chromosome 22 was found to contain at least one NFI sequence in the functional class. The distribution of these sites across the human chromosome (Figure 4), however, is nonuniform, and appears to correlate with the distribution of known and predicted genes ($R = 0.65$ for the functional and 0.71 for the nonrandom class of NFIs to the combined annotated and GENSCAN predicted genes). Thus, there is a distinct distribution of the eukaryotic transcription factor in the human sequence, as one would expect. Furthermore, we see that it is not simply the occurrence of binding sites, but their distribution that define the function of NFI as a transcriptional activator.

Using a value for $P_{\mathrm{Zmin}} = 0.5$ kb, we have located 7580 ZDRs in human chromosome 22, representing 0.454% of the known sequence (Table 1), or that there is one ZDR in ~4462 bp. Overall, the number of ZDRs in chromosome 22 is lower than predicted from random occurrence (for a $P_{\mathrm{Zmin}} = 0.5$ kb) by >6-fold. The distribution of observed ZDRs into genes, pseudogenes and predicted genes are essentially what one would expect from random occurrence when considering the percent of the chromosome sequence represented by each sequence type (Table 1). In contrast, the number of genes, pseudogenes and predicted genes that contain ZDRs differ, with nearly all known genes (96.6%) and predicted genes (87%) having at least one ZDR, while fewer than half (42.7%) of the pseudogenes contain ZDRs. It is interesting at this point to see that the number of gene sequences with ZDRs at or near the TSS is ~77% of the total, while those with ZDRs at the 3′ end is only approximately one-third (33.4%). This discrepancy between the two ends of the sequences is most dramatic for the annotated genes, but is also seen for pseudogenes and predicted genes. Thus, unlike the NFIs, ZDRs

Table 2. Occurrence of NFI binding sites predicted in human chromosome 22 (16,17)

Sequence type	Number of sequences w/NFIs (%)			Total number of NFIs (predicted[a])		
	Total ($P_{NFI} \geq 500$)	Nonrandom[b] ($500 \geq P_{NFI} \geq 5000$)	Strong[b] ($P_{NFI} \geq 5000$)	Total ($P_{NFI} \geq 500$)	Nonrandom[b] ($500 \geq P_{NFI} \geq 5000$)	Strong[b] ($P_{NFI} \geq 5000$)
Chromosome 22				90777 (70 905)	80959 (64 680)	9818 (6225)
Genes	383 (100%)	383 (100%)	373 (97.4%)	45619[b] (34 461)	40597[b] (31 461)	5022[b] (3025)
Pseudogenes	234 (100%)	233 (97.6%)	139 (59.4%)	3514[b] (2737)	3121[b] (2497)	393[b] (240)
Predicted genes	790 (100%)	788 (99.8%)	730 (92.4%)	67371[b] (51 295)	59845[b] (46 792)	7526[b] (4504)

[a]Numbers of NFI binding sites predicted for genes, pseudogenes and predicted genes are based on the actual number of sites observed multiplied by the fraction of the total number of base pairs represented by that sequence type.
[b]NFI sites that appear in two different overlapping gene, pseudogene and/or predicted gene sequences are counted for both sequences.

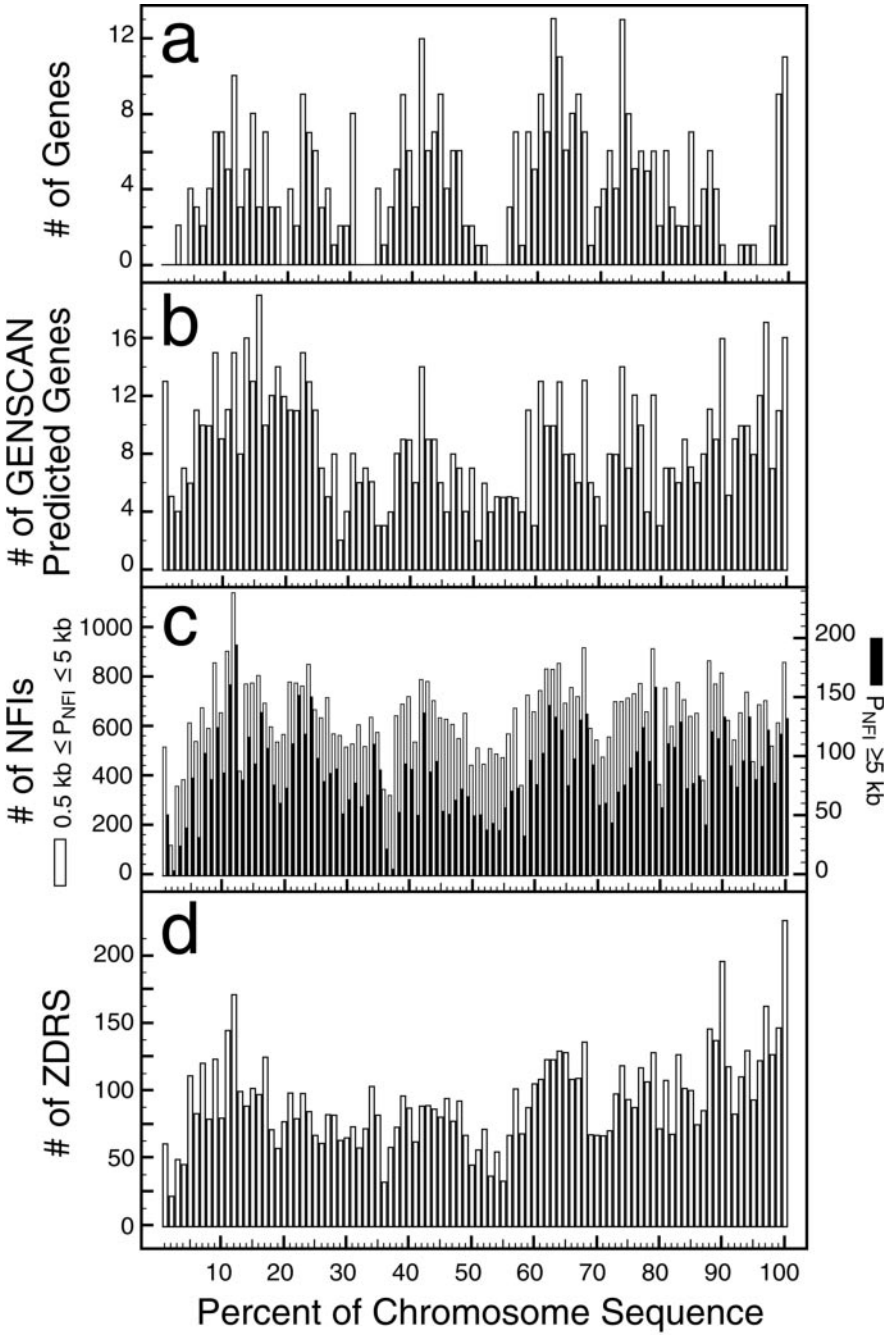

Figure 4. Distribution of genes (**a**), GENSCAN predicted genes (**b**), NFI sites (**c**) and ZDRs (**d**) along the human chromosome 22 sequence. The occurrence of each sequence type is summed for each 1% increment of the chromosomal sequence.

are much less frequent and, when they do occur, tend to be more discriminate in location.

The distribution of ZDRs across chromosome 22 very closely mirrors that of NFIs (Figure 4d), with correlation between the number of nonrandom NFIs and ZDRs across the chromosome giving an *R*-value of 0.72 (Figure 5). There is no a priori reason why NFIs and ZDRs should be correlated. Sequences that are strong NFIs do not show an alternating purine–pyrimidine pattern and are not necessarily rich in C·G base pairs; therefore, they should be very weak ZDRs (sequences in the functional class of NFI sites, with $P_{NFI} \geqslant 5$ kb, have P_Z at least one order of magnitude lower than the P_{Zmin} of 0.5 kb used to identify ZDRs). The correlation between ZDRs and the predicted and known genes across the chromosome ($R = 0.59$) is not as strong as that seen for NFIs. However, applying a Spearman Rank Correlation test indicates that the probability of having no correlation between the distributions of ZDRs and the known genes is <7 in 10 000 (p-value = 0.0007), and <1 in 10 000 (p-value = 0.0001) against the predicted genes.

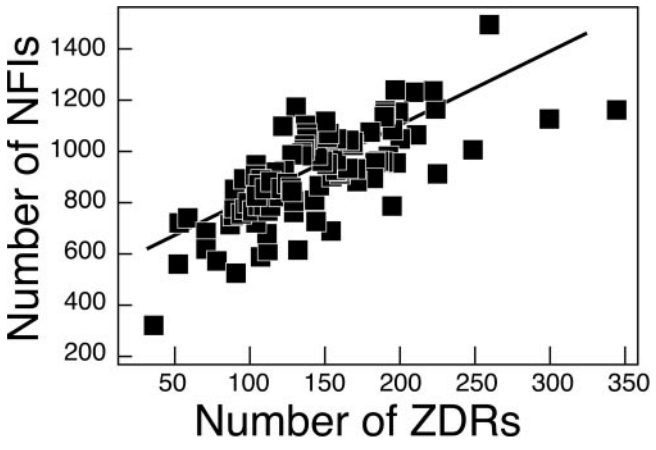

Figure 5. Correlation between the distribution of NFI and ZDR sites across human chromosome 22. The values from Figure 4 for NFIs (4c) and ZDRs (4d) are plotted against each other. The line is a linear least squares fit with $R = 0.72$.

Distribution of NFIs and ZDRs around the gene termini

Both NFI and ZDR sequences accumulate around the known TSS of the human genes. The number of nonrandom NFIs ($P_{NFI} \geqslant 0.5$ kb) rises above background levels from ~1 kb upstream to ~1 kb downstream of the TSS (a similar, but less pronounced distribution is seen for the functional class of NFIs) and has a maximum count at ~140 bp downstream of the TSS (Figure 6). At the 3′ terminus of these known genes, the number of NFIs shows a very distinct dip. Thus, the distributions of NFIs across the known genes in chromosome 22 correlate with the known function of these sites, accumulating around the 5′ end, and suppressed where specific sequence signals are required for transcription termination.

We had previously seen in a more limited study (11) that ZDRs accumulate at the TSS of human genes. In the current study, we observe that 295 of the 383 identified genes (or >77%) in chromosome 22 have ZDRs within ±1 kb of the TSS. The larger sample size in the current study reveals a defined distribution of ZDRs centered upstream of the TSS by ~40 bp (and upstream of the NFI distribution by ~180 bp) and extending ~1 kb in either direction (Figure 6b). At the 3′ termini, ZDRs are distributed apparently randomly, but we should note that the very small number of ZDRs in this region does not allow us to determine whether such sequences are suppressed at this end. Thus, we see a strong correlation between the distribution of ZDRs and NFIs across human genes, with ZDRs accumulating upstream of both NFI and TSS loci.

This accumulation of NFIs and ZDRs around the TSS of the human genes follows the distributions of C/G base pairs (%C + G content, Figure 6c) and d(CpG) dinucleotides (data not shown). One can argue that the accumulation of the ZDRs is the direct consequence of the C/G content around the TSS. However, the majority of the ZDRs in chromosome 22 are CA/TG type sequences (which are only 50% C + G). In addition, we see that C + G nucleotides are distributed broadly both upstream and downstream of the TSS, while ZDRs are biased upstream and NFIs downstream of the TSS. Thus, we suggest that it is not simply the C + G content that determines the accumulation of ZDRs around the TSS. Indeed, it is possible to argue the opposite point that the C + G distribution is a consequence of the distribution of ZDR and NFI (and potentially other high C/G content transcriptional elements), given that the C + G distribution appears to mirror the sum of the ZDR and NFI distributions, but neither in themselves. For example, although we would expect TATA promoters to be favored in T + A rich regions upstream of a TSS, we would conclude that it is the localization of the promoters that defines the distribution of the nucleotide content and not vice versa. In support of this latter point for the current study, we see that the residual difference between the normalized ZDR and C + G distributions around the TSS of genes is a broad distribution with a maximum ~200 bp downstream of the TSS, similar overall to the distribution of NFI sites.

With the strong preference of Z-DNA for alternating CG dinucleotides, one can ask whether the occurrence of ZDRs seen here correlates with CpG islands. We observed that with a $P_{Zmin} \geqslant 0.5$ kb, only 9.3% of the ZDRs in chromosome 22 fell into known CpG islands (identified at ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/maps/mapview/BUILD.34.3/). A higher percentage of ZDRs located within 1 kb of the TSS are within CpG islands (28.2%), reflecting the preference of both motifs upstream of genes. However, it is clear from this analysis that the ZDRs identified in this study do not simply mirror CpG islands. This is consistent with the large number of ZDRs that are primarily alternating CA/TG repeats.

Similarly, only 9.2% of the identified NFI binding sites across human chromosome 22 fall into known CpG islands. Again, this increases to 28.3% for NFI sites identified ±1 kb of the TSS of known genes. It is clear, therefore, that neither the ZDR and NFI sites identified here are the direct result of the localization of CpG islands upstream of the initiation sites for transcription in these genes.

Interestingly, the distributions of NFIs and ZDRs across predicted genes mirror that seen in the known genes, but only for those sequences predicted by GENSCAN to have introns (data not shown). These distributions are significantly suppressed, or flat for GENSCAN sequences that lack introns. Finally, the analysis of *E.coli* genes shows that ZDRs and NFIs

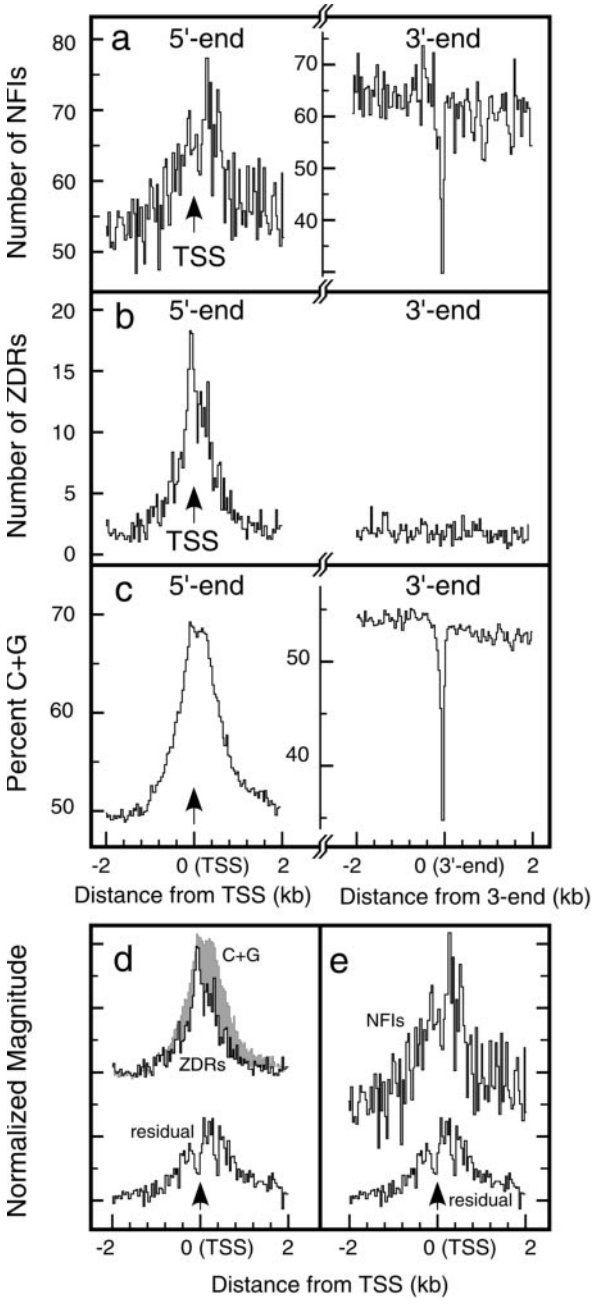

Figure 6. Distribution of NFIs (**a**), ZDRs (**b**), and the C + G content (**c**) around the transcription start sites (TSS) and 3′ terminus of known genes in human chromosome 22. The genes in the chromosomal sequence are aligned with their TSS or 3′ ends (arrows), and the number of each sequence type are summed in 200 bp increments, starting 2 kb upstream and extending 2 kb downstream of both the TSS and 3′ ends of the aligned genes. Values for the 3′ ends for the NFI and percent C + G are offset from the TSS distributions to fit into the panels. (**d**) Comparison of numbers of ZDRs with C + G content around the TSS of human genes. The upper figure shows an overlay of (b) and (c), assuming that a 1% increase in C + G content is associated with an increased probability of seeing one ZDR. The lower panel shows the residual resulting from subtracting the two distributions. (**e**) Comparison of the residual from (d), bottom figure, to the distribution of potential NFIs (a) around the human TSS, top figure.

are strongly suppressed at both the TSS and the 3′ end, reflecting the incompatibility of well-defined transcription promoter, initiation and termination sites in prokaryotic genes with both the eukaryotic promoter and ZDRs (Figure 7).

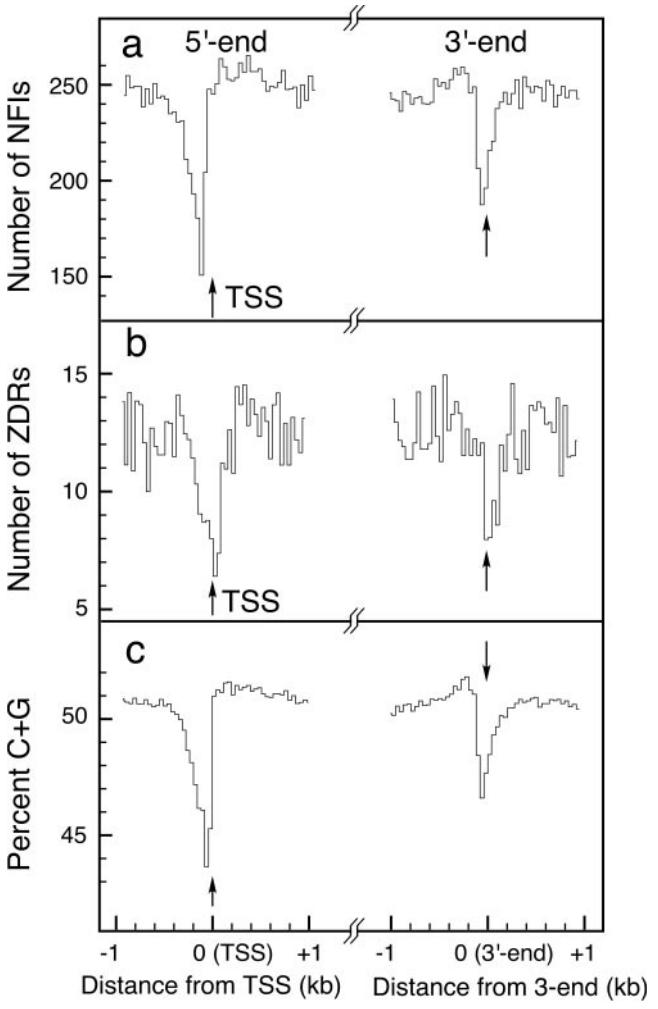

Figure 7. Distribution of NFIs (**a**), ZDRs (**b**) and the C + G content (**c**) around the transcription start sites (TSS) and 3′ terminus of genes in the *E.coli* genome. The genes are aligned with their TSS or 3′ ends (arrows), and the number of each sequence type are summed in 200 bp increments, starting 1 kb upstream and extending 1 kb downstream of both the TSS and 3′ ends of the aligned genes.

DISCUSSION

In this study, we have applied an *in silico* approach to identify potential Z-DNA regions (ZDRs) and binding sites for the NFI transcription factor in the human chromosome 22 sequence (16,17). The distribution of ZDRs across the genomic sequence and at the transcriptional start sites of genes, remarkably, is nearly identical to those of NFI sites; thus, if we were to classify regulatory elements according to their genomic distributions, the behavior of ZDRs could be considered to be similar to that of a eukaryotic nuclear factor binding site. The primary question that we addressed in the current study is whether the mechanism of Z-DNA coupled transcription regulation, as characterized in the CSF-1 gene by Liu *et al*. (5), is common to a set of human genes. In this mechanism, left-handed Z-DNA serves to maintain a gene in its activated state after initial activation by the BAF and NFI complexes. It has been shown that Z-DNA does not bind to nucleosomal proteins (12–15) and, therefore, would inhibit the reestablishment of a

Table 3. Number of genes with strong ($P_{NFI} \geqslant 5$ kb) NFI binding sites that also have ZDRs, both within 1 kb of their transcriptional start sites (TSS)

BAF regulated genes [37 annotated of 82 total (5)]					Chromosome 22 genes [383 annotated (16,17)]				
Total	ZDR upstream (bp)		ZDR downstream (bp)		Total	ZDR upstream (bp)		ZDR downstream (bp)	
11	⩽150	3	⩽150	3	156	⩽150	26	⩽150	36
	150–300	1	150–300	1		150–300	17	150–300	17
	⩾300	2	⩾300	1		⩾300	27	⩾300	33

Listed are the total number of sequences with NFIs and ZDRs in the BAF regulated genes (5) and identified in human chromosome 22, along with the number of sequences with ZDRs that are ⩽150 bp, between 150 and 300 bp, and ⩾300 bp upstream or downstream of the strongest NFIs in each group.

stable chromatin structure. The formation of Z-DNA would be induced by the negative superhelices generated behind an active RNA polymerase, as originally proposed in the twin-domain model (22). This is consistent with the observation made previously (11), and more definitively demonstrated here, that Z-DNA elements accumulate near the TSS of human genes. In addition, it has been shown that the Z-DNA binding domain from ADAR1 can in itself act as a *cis*-regulatory element in yeast (23).

For the current analysis, we first characterized the ZDR and NFI sites for the sequences identified as being affected by the BAF complex (5). Of the 80 mRNAs that were observed to be BAF regulated, only 37 are associated with fully annotated gene sequences at the time of this study. All but one of these known genes contained at least one functional NFI site with $P_{NFI} \geqslant 5$ kb and within ±1 kb of the TSS. This is not surprising since the NFI promoter is expected to be one of the common elements in these BAF regulated genes. Of these, 11 (≈30% of the 37) also have ZDRs within 1 kb of the TSS (Table 3). We can conclude therefore that nearly one-third of the genes associated with BAF regulation may potentially involve Z–DNA in a manner analogous to that of the CSF-1 gene.

For the annotated gene sequences in human chromosome 22, 370 (∼97%) have nonrandom NFI sites and 295 have ZDRs within ±1 kb of their TSS (Table 3). Of these, 158 sequences contain both potentially functional NFI and ZDR sites within 1 kb of the TSS. Thus, if the mechanism for Z-DNA and NFI activation in the BAF regulated genes is general, we would predict that ∼41% of the genes in human chromosome 22 could potentially be similarly regulated.

The model that couples Z-DNA to BAF-regulation of human genes as proposed by Liu *et al*. (5), however, may impose additional constraints to the architecture of promoter elements within this class of gene sequences. For example, in order to maintain an activated gene-state, there may be the additional requirement that Z-DNA elements also be upstream of the NFI binding sites so that the spatial orientation of the promoter and TSS are not adversely affected by the transcriptionally induced left-handed double helix. It may also be necessary to place the Z-DNA elements within one or two nucleosome distances (∼100–300 bp) from the NFI site so that it is the nucleosome structure at or near the activation site that is affected by the structural transition. For the gene sequences identified by Liu *et al*. (5) and in chromosome 22, near equal numbers of sequences have their closest ZDRs upstream and downstream of the strongest NFIs (Table 3). However, ZDRs associated with NFI sites that are upstream of the TSS are also twice as likely to be upstream of the NFI site. Similarly, ZDRs are twice as likely to be downstream of the NFI site if the promoter sequence is also downstream of the TSS. Thus, under these constraints, we would predict that 44 genes (∼11%) of the annotated genes in chromosome 22 may be regulated by a mechanism similar to that for CSF-1, with their ZDRs located within 300 nt upstream of a strong NFI site. The large number of sequences that have Z-DNA sites downstream of the NFI sites suggests a separate class of genes that may be regulated through a variation on this coupled mechanism, with the transcriptionally driven formation of Z-DNA serving to attenuate the rates of gene expression (11). The possibility of this negative mechanism of gene regulation by the Z-form of DNA has yet to be experimentally explored in detail. However, it has been shown that Z-DNA placed near the TSS of some genes will inhibit rather than enhance gene expression (24,25).

The mechanism of Z-DNA coupled transcriptional activation may be general and extend beyond the NFI class of promoters. For example, Wang *et al*. (26) reported that a Z-DNA forming segment at the proximal core promoter of the mouse MARCKS promoter is important for basal transcription. Thus, the regulation of transcription by eukaryotic promoters is not dependent solely on the binding properties of the protein to the cognate DNA, but the dynamic nature of the DNA double-helix now appears to play an important role in maintaining an activated state. Obviously, the predictions for such regulation of specific genes need to be tested experimentally to confirm their mechanism of regulation. For example, formation of Z-DNA within these upstream sequences would require release of negative superhelical density resulting from dramatic remodeling of the chromatin structure (which may be the link to BAF regulation). In the absence of this superhelical energy, the potential ZDRs would likely remain in their standard right-handed conformation and, therefore, would not participate in the regulatory mechanism. Thus, additional work must be done to determine whether such topological strain can be induced for the individual genes. In addition, we recognize that other accessory protein factors would need to be incorporated into the thermodynamic model in order to increase our confidence in identifying functional promoters at the potential NFI sites identified here. This can be explored further as we survey genomic sequences for other transcription-factor binding sites and correlate them to various architectural DNA elements such as Z-DNA.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Clifford Pereria in the Statistics Facility Core of the Environmental Health Sciences Center (EHSC) at Oregon State University for help in the statistical analysis. This work was funded by grants from the National Institutes of Health (R1GM62957A) and the National Science Foundation (MCB0090615) to P.S.H., and by the Environmental Health Sciences Center at Oregon State University (NIEHS ES00210).

REFERENCES

1. Wang,A.H., Quigley,G.J., Kolpak,F.J., Crawford,J.L., van Boom,J.H., van der Marel,G. and Rich,A. (1979) Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature*, **282**, 680–686.
2. Marx,J. (1985) Z-DNA: still searching for a function. *Science*, **230**, 794–796.
3. Herbert,A., Lowenhaupt,K., Spitzner,J. and Rich,A. (1995) Double-stranded RNA adenosine deaminase binds Z-DNA *in vitro*. *Nucleic Acids Symp. Ser.*, **33**, 16–19.
4. Schwartz,T., Rould,M.A., Lowenhaupt,K., Herbert,A. and Rich,A. (1999) Crystal structure of the Zalpha domain of the human editing enzyme ADAR1 bound to left-handed Z-DNA. *Science*, **284**, 1841–1845.
5. Liu,R., Liu,H., Chen,X., Kirby,M., Brown,P.O. and Zhao,K. (2001) Regulation of CSF1 promoter by the SWI/SNF-like BAF complex. *Cell*, **106**, 309–318.
6. Herbert,A. and Rich,A. (1999) Left-handed Z-DNA: structure and function. *Genetica*, **106**, 37–47.
7. Herbert,A., Schade,M., Lowenhaupt,K., Alfken,J., Schwartz,T., Shlyakhtenko,L.S., Lyubchenko,Y.L. and Rich,A. (1998) The Zalpha domain from human ADAR1 binds to the Z-DNA conformer of many different sequences. *Nucleic Acids Res.*, **26**, 3486–3493.
8. Kim,Y.G., Lowenhaupt,K., Maas,S., Herbert,A., Schwartz,T. and Rich,A. (2000) The zab domain of the human RNA editing enzyme ADAR1 recognizes Z-DNA when surrounded by B-DNA. *J. Biol. Chem.*, **275**, 26828–26833.
9. Rahmouni,A.R. and Wells,R.D. (1989) Stabilization of Z DNA *in vivo* by localized supercoiling. *Science*, **246**, 358–363.
10. Lukomski,S. and Wells,R.D. (1994) Left-handed Z-DNA and *in vivo* supercoil density in the *Escherichia coli* chromosome. *Proc. Natl Acad. Sci. USA*, **91**, 9980–9984.
11. Schroth,G.P., Chou,P.J. and Ho,P.S. (1992) Mapping Z-DNA in the human genome. Computer-aided mapping reveals a nonrandom distribution of potential Z-DNA-forming sequences in human genes. *J. Biol. Chem.*, **267**, 11846–11855.
12. Nickol,J., Behe,M. and Felsenfeld,G. (1982) Effect of the B–Z transition in poly(dG-m5dC).poly(dG-m5dC) on nucleosome formation. *Proc. Natl Acad. Sci. USA*, **79**, 1771–1775.
13. Ausio,J., Zhou,G. and van Holde,K. (1987) A reexamination of the reported B–Z DNA transition in nucleosomes reconstituted with poly(dG-m5dC).poly(dG-m5dC). *Biochemistry*, **26**, 5595–5599.
14. Casasnovas,J.M. and Azorin,F. (1987) Supercoiled induced transition to the Z-DNA conformation affects the ability of a d(CG/GC)12 sequence to be organized into nucleosome-cores. *Nucleic Acids Res.*, **15**, 8899–8918.
15. Garner,M.M. and Felsenfeld,G. (1987) Effect of Z-DNA on nucleosome placement. *J. Mol. Biol.*, **196**, 581–590.
16. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
17. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
18. Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
19. Ho,P.S., Ellison,M.J., Quigley,G.J. and Rich,A. (1986) A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. *EMBO J.*, **5**, 2737–2744.
20. Johnston,B.H., Ohara,W. and Rich,A. (1988) Stochastic distribution of a short region of Z-DNA within a long repeated sequence in negatively supercoiled plasmids. *J. Biol. Chem.*, **263**, 4512–4515.
21. Roulet,E., Bucher,P., Schneider,R., Wingender,E., Dusserre,Y., Werner,T. and Mermod,N. (2000) Experimental analysis and computer prediction of CTF/NFI transcription factor DNA binding sites. *J. Mol. Biol.*, **297**, 833–848.
22. Liu,L.F. and Wang,J.C. (1987) Supercoiling of the DNA template during transcription. *Proc. Natl Acad. Sci. USA*, **84**, 7024–7027.
23. Oh,D.B., Kim,Y.G. and Rich,A. (2002) Z-DNA-binding proteins can act as potent effectors of gene expression *in vivo*. *Proc. Natl Acad. Sci. USA*, **99**, 16666–16671.
24. Sheridan,S.D., Benham,C.J. and Hatfield,G.W. (1999) Inhibition of DNA supercoiling-dependent transcriptional activation by a distant B-DNA to Z-DNA transition. *J. Biol. Chem.*, **274**, 8169–8174.
25. Sheridan,S.D., Opel,M.L. and Hatfield,G.W. (2001) Activation and repression of transcription initiation by a distant DNA structural transition. *Mol. Microbiol.*, **40**, 684–690.
26. Wang,L., Liu,X. and Lenox,R.H. (2002) Transcriptional regulation of mouse MARCKS promoter in immortalized hippocampal cells. *Biochem. Biophys. Res. Commun.*, **292**, 969–979.