THE JOURNAL OF BIOLOGICAL CHEMISTRY
© 1992 by The American Society for Biochemistry and Molecular Biology, Inc.

Vol. 267, No. 17, Issue of June 15, pp. 11846-11855, 1992
Printed in U.S.A.

Mapping Z-DNA in the Human Genome

COMPUTER-AIDED MAPPING REVEALS A NONRANDOM DISTRIBUTION OF POTENTIAL Z-DNA-FORMING SEQUENCES IN HUMAN GENES*

(Received for publication, September 16, 1991)

Gary P. Schroth, Ping-Jung Chou‡, and P. Shing Ho§
From the Department of Biochemistry and Biophysics, Oregon State University, Corvallis, Oregon 97331

In this work, we have predicted and mapped the potential Z-DNA-forming sequences in over one million base pairs of human DNA, containing 137 complete genes. The computer program (Z-Hunt-II) developed for this study uses a rigorous thermodynamic search strategy to map the occurrence of left-handed Z-DNA in genomic sequences. The search algorithm has been optimized to search large sequences for the potential occurrence of Z-DNA, taking into account sequence type, length, and cooperativity for a given stretch of potential Z-DNA-forming nucleotides. In this extensive data set we have identified 329 potential Z-DNA-forming sequences. The exact locations of the potential Z-DNA-forming sequences in the data set have been mapped with respect to the location of structural features of the genes. This analysis reveals a distinctly nonrandom distribution of potential Z-DNA-forming sequences across human genes and, most notably, that strong Z-DNA-forming sequences are more commonly found near the 5′ ends of genes. We find that 35% of the Z-DNA-forming sequences are located upstream of the first expressed exon, while only 3% of the sequences are located downstream of the last expressed exon. The remaining 62% of the Z-DNA-forming sequences, which are located either in introns (47.1%) or exons (14.9%), are also nonrandomly distributed, with a strong bias toward locations near the site of transcription initiation. We interpret this distribution of potential Z-DNA-forming sequences toward the 5′ end of human genes in terms of the well established "twin-domain model" of transcription-induced supercoiling and the effect of this topological strain on Z-DNA formation in eukaryotic cells.

Within the next decade we will be obtaining new and increasingly large amounts of DNA sequence information, perhaps even the sequence of the entire human genome, through the efforts of the combined genome projects (Maddox, 1991). It has been suggested that such a large increase in DNA sequence data may affect the course of experimental molecular biology, requiring that we as research biologists ask more interesting questions of the information in this data base (Gilbert, 1991). Even now with the currently available sequence data, one of the major challenges in molecular biology is the prediction and mapping of relevant "signals" in a given DNA sequence, based upon biological and biophysical principles. These include experimentally determined biochemical regulatory "signals" such as consensus binding sites for sequence-specific DNA binding proteins. Local structural "signals" however, are also determined by the base sequence of a DNA molecule. Ideally, armed with an understanding of the relationship between DNA base sequence and DNA structure, one would like to be able to predict local structural features of specific regions of DNA and, in this way gain further insight into the relation between structure and function in genomic processes. In this paper, we use a computer-aided, thermodynamic search strategy to predict and map potential left-handed Z-DNA-forming sequences in a large data set of human DNA.

It is now well established that DNA structure is polymorphic, and that many sequence-specific non-B-DNA conformations exist, often times in response to changes in the environmental conditions (Wells, 1988; Kennard and Hunter, 1989). Within the past several years much progress has been made into the structural and chemical aspects of many non-B-DNA conformations, however the biological relevance of any of the "unusual" DNA structures (Wells, 1988) has still not been well established. One of the more dramatic structural transitions observed in DNA is that between right-handed B- and left-handed Z-DNA (Rich *et al.*, 1984). Ever since the structure of Z-DNA was first solved by x-ray diffraction (Wang *et al.*, 1979), Z-DNA has been under intense investigation and now stands as arguably the best understood of all non-B-DNA conformations. The local flipping of small regions of B-DNA to Z-DNA in topologically constrained DNA molecules requires negative supercoiling and is strongly favored in alternating purine/pyrimidine (APP)[1] sequences (reviewed by Rich *et al.* (1984) and Jovin *et al.* (1987)). APP sequences allow the nucleotides to assume their lowest energy conformation as Z-DNA, with purines in the *syn*, and pyrimidines in the *anti* conformation. Because repeated purine/pyrimidine sequences are strongly favored in Z-DNA, it is useful to think of the dinucleotide as the fundamental repeating unit of Z-DNA (Jovin *et al.*, 1987). As such, there exists a hierarchy in the ability of naturally occurring dinucleotides to form Z-DNA, where GC is more favored than (GT)/(AC), which are strongly favored over AT (Kagawa *et al.*, 1989). In fact, A/T base pairs dramatically inhibit Z-DNA formation, and long stretches of alternating AT will not form Z-DNA,

* This work was supported in part by American Cancer Society Grant NP-740 (to P. S. H.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ Also supported by Institute of General Medical Sciences Grant GM-43133 (to Dr. W. C. Johnson, Oregon State University).

§ Recipient of a Junior Faculty Research Award 306 from the American Cancer Society. To whom correspondence should be addressed: Dept. of Biochemistry and Biophysics, 535 Weniger Hall, Oregon State University, Corvallis, OR 97331. Tel.: 503-737-2769; Fax: 503-737-0481.

[1] The abbreviations used are: APP, alternating purine/pyrimidine; bp, base pairs; TSS, transcription start site; UTR, untranslated region.

even under conditions of high negative supercoiling (Panyutin *et al.*, 1985; McClellan *et al.*, 1986).

A number of studies have utilized these simple APP sequence rules to search for the potential occurrence of Z-DNA in genomic sequences (Konopka *et al.*, 1985; Trifonov *et al.*, 1985; Braaten *et al.*, 1988; Rollo *et al.*, 1989). In short, these searches find that APP and particularly potential Z-DNA sequences as defined by the APP criteria are underrepresented in genomic sequences. These studies, however, fall short in two respects. First, the types of sequences studied represent only a small part of the available data base and do not generally represent a homogeneous set of sequence data. Thus, it is difficult to generalize trends for the potential occurrence of Z-DNA and draw conclusions concerning the possible biological function of this structure. Second, and perhaps more important, the formation of Z-DNA is not restricted to only APP sequences. Biochemical studies have shown that a large number of perturbations to the APP rule can be accommodated by the left-handed structure, including placing pyrimidines in the disfavored *syn* conformation, or changing the phase of the alternation of bases (Ellison *et al.*, 1985; McLean *et al.*, 1988; McLean and Wells, 1988). When these additional sequence parameters are taken into account, the number of possible Z-DNA-forming sequences that can occur in a genome increases dramatically. The question then becomes how does one develop a search strategy to account for these variations to the APP rule and how can we assess the relative ability of these sequences to form Z-DNA (*i.e.* rank the "Z-ness" of the sequences) found by this strategy? Our approach has been to use the thermodynamic propensity for each dinucleotide combination to assess the ability of a particular sequence to adopt the Z conformation.

Currently the B- to Z-DNA transition energy for all the possible dinucleotide combinations in DNA have been either directly measured or derived from the behavior of cloned sequences in closed circular, negatively supercoiled plasmids (summarized in Ho *et al.*, 1986). The experimentally determined B- to Z-DNA transition energies had previously been incorporated into a computer program, called Z-Hunt, which uses these energies to analyze DNA sequences according to their thermodynamic propensity to form Z-DNA (Ho *et al.*, 1986). The analysis of ϕX-174, pBR322, and SV40 DNA sequences with Z-Hunt properly predicted the majority of Z-DNA sites located from Z-DNA-specific antibody studies on these genomes. In addition, the analysis in most cases properly predicted the relative propensity for these sequences to form Z-DNA as measured by the relative probability for antibody binding at these sites.

In this study, we have mapped potential Z-DNA-forming sequences across 137 different human genes using this thermodynamic approach. Our most recent version of this algorithm (Z-Hunt-II) has been optimized to analyze large genomic sequences according to these thermodynamic criteria. This is the first study which maps Z-DNA, or any other non-B-DNA structure, in such a large homogeneous data set (containing over one million base pairs of human DNA). We show that there is a distinctly nonrandom distribution of potential Z-DNA-forming sequences in the human genome. In the genes studied, 35% of the potential Z-DNA-forming sequences are located upstream of the first expressed exon, while only 3% are found downstream of the last expressed exon. This raises some interesting consequences in light of the twin-domain model of Liu and Wang (1987) which proposes that, during transcription, the topology of local DNA domains are dynamic, with positive supercoils accumulating in front of the transcriptional apparatus and negative supercoils accumulating behind. This model suggests that the location of many of these sequences, especially those located near the 5′ end of the transcription unit, are likely to be in a dynamic negatively supercoiled environment during transcription of the gene, which could potentially drive the formation of Z-DNA. Mapping the location of strong Z-DNA-forming regions near specific genes is thus an important step in ascertaining any potential function Z-DNA may have in transcriptional regulation.

MATERIALS AND METHODS

The program used for mapping Z-DNA in large genomes was developed by extending the basic strategy from the original thermodynamic search strategy of Z-Hunt, as described by Ho *et al.* (1986). In short, the search strategy relies on our ability to predict properties of the B- to Z-DNA transition induced by negative supercoiling in closed circular DNA. The program walks along the length of a genomic sequence in fixed search windows. The nucleotide sequence within each window is then placed in the context of a theoretical 5000-bp closed circular plasmid, and the probability for inducing Z-DNA formation within the insert is calculated using a statistical mechanics treatment of the zipper model for the B- to Z-DNA transition (Peck and Wang, 1983). The propensity of the insert to adopt the Z conformation is determined, in the original Z-Hunt program, as the superhelical density at which one base pair is induced to form Z-DNA, representing the onset of the B- to Z-DNA transition, within the insert. In the current program, this is defined at the point in the B- to Z-DNA transition having the maximum slope, *i.e.* where change in twist *versus* the change in superhelical density is at a maximum. Since the transition is cooperative, this would be analogous to defining the affinity of a cooperative protein, such as hemoglobin, for its substrate at the midpoint of the binding curve; thus we now consider the cooperative mechanism inherent in the structural transition when assessing the ability of a sequence to form Z-DNA.

A comparison of the calculated superhelical density to the average superhelical density calculated for Z-DNA formation in a 50,000-bp randomly generated sequence, once corrected for the window size, defines the Z-Score. Thus in practical terms, the Z-Score relates the ability for a given sequence to adopt the Z conformation relative to a random sequence. We can also interpret the Z-Score as the number of random sequences that must be searched to find a nucleotide sequence that is as good or better at forming Z-DNA than the sequence in question. This definition in the original Z-Hunt program was shown to correlate to the affinity of anti-Z-DNA antibodies to specific Z-DNA sites in various genomes and, thus, can be related to an experimentally determined measure for the ability of a sequence to form Z-DNA (Ho *et al.*, 1986).

The resulting program for the current studies (Z-Hunt-II) includes a number of modifications over the original Z-Hunt algorithm. First, as discussed above, the cooperativity of the B- to Z-DNA transition has been incorporated into the measure for the probability of Z-DNA formation. In addition, the optimum length (within a set range) for Z-DNA formation within a fragment is now calculated, as opposed to simply the probability for Z-DNA formation in fixed-length fragments. Finally, the resolution of the calculation has been improved in that a Z-Score is now calculated for each base pair as opposed to each half turn (6 bp) of Z-DNA. To accommodate these changes, and still allow the program to search large genomic sequences, the strategy for finding the superhelical density used to calculate the Z-Score has been modified. In the original Z-Hunt program, the superhelical density at the onset of the transition was determined by incrementally increasing the negative supercoils in the 5000-bp plasmid until formation of one base pair as Z-DNA was induced in the sequence of interest. In the current version, Z-Hunt-II, we start at two extremes in superhelical density (both above and below the midpoint of the transition) and converge toward the point having the maximum slope in the transition. This effectively reduces the number of statistical mechanics calculations by a factor of 10 for an average sequence. The resulting Z-Scores from Z-Hunt-II differ from those values obtained from Z-Hunt by approximately one order of magnitude, reflecting the additional cooperativity and length information of the new algorithm. The relative magnitude of the Z-Scores for any given sequence, however, still reflects the propensity for that sequence to adopt the Z conformation compared to other DNA sequences.

Briefly, the search strategy within Z-Hunt-II works through the following steps: 1) a search window that will be used to walk along a

DNA sequence is defined; 2) each nucleotide of the sequence within the search window is assigned its energetically most favored base conformation (either *anti* or *syn*) in the context of the entire sequence in the search window; 3) the free energy associated with each nucleotide in this base conformation is assigned according to the base conformations; 4) the search window is placed within a theoretical 5000-bp closed circular plasmid (analogous to the actual experimental system used to measure the stability of Z-DNA); 5) the superhelical density at the midpoint of the supercoil induced B- to Z-DNA transition for the sequence is calculated; 6) the "Z-Score" of the sequence is calculated, and is defined as the probability of finding a random sequence that is as good or better at forming Z-DNA as that in the search window; and finally, 7) this is repeated as the window walks one nucleotide at a time along the entire sequence.

Z-Hunt-II has been written in the C programming language to run on an IBM PC-XT or AT, or compatible, microcomputers with or without a math coprocessor. Our analysis was performed on a Hyundai Super-286C computer (an IBM-AT compatible with a math coprocessor), which is capable of analyzing approximately 100 bp per min.

RESULTS

The elucidation of the biological role of non-B forms of DNA may help us in understanding the many possible mechanisms by which genetic processes are regulated. One approach toward this end would be to map the potential occurrence of these structures in a large homogeneous set of nucleic acid sequence and ask whether the distribution of conformations correlate to the positions of biological functions. The present studies map the potential Z-DNA-forming sequences in human genes using a thermodynamic rather than a sequence-matching search strategy. To undertake this task, we needed to first develop a homogeneous data set from all available sequences of the current human genome and second, optimize our search algorithm to accommodate such a large data set.

The Z-Hunt-II Program—Z-Hunt-II is a program developed to search for the occurrence of Z-DNA in genomic sequences using a rigorous thermodynamic search strategy. The program does not simply consider the stability of Z-DNA as a difference in free energy between the right-handed B- and left-handed Z-DNA conformations of DNA (ΔG), but utilizes a statistical mechanical treatment of negative supercoiling-induced Z-DNA formation (Ho *et al.*, 1986). This strategy takes into account effects of differences in base sequence, length, and cooperativity for a stretch of potential Z-DNA-forming nucleotides. This is then translated into a quantitative measure, the Z-Score, which is the probability that Z-DNA will form in that sequence within the context of the whole genome. A good Z-DNA-forming sequence will have a high Z-Score, and therefore would require less negative supercoiling to form Z-DNA compared to a sequence with a lower Z-Score. This definition of the Z-Score, to assess the probability that Z-DNA will form in a sequence, has previously been shown to correspond well with the probability observed for binding anti-Z-DNA antibodies to sequences in the ϕX-174 genome (Ho *et al.*, 1986).

Since the Z-Score is dependent upon length, we have chosen to analyze sequences with a fixed range of "window" sizes from 6 to 8 dinucleotides, equal to 12–16 bp of DNA or 1–1.3 turns of Z-DNA. This particular window size was chosen for two reasons: 1) the lower limit of 12 bp, or one full turn of Z-DNA, is generally accepted as a reasonable minimum length to form a left-handed Z-DNA helix in the context of B-DNA, from both *in vitro* and *in vivo* studies (in some instances stretches of DNA helix as short as 8 bp have been reported to form Z-DNA (Nordheim and Rich, 1983a)); and 2) in terms of the Z-Hunt-II program, a setting of 6–8 dinucleotides seems to best balance search speed with accuracy. It is worth noting that one would get different results depending upon the selection of window size, since the length of the potential Z-DNA-forming domain is very important in determining the Z-Score. Potential Z-DNA sequences that extend beyond the set window size would appear in this analysis as contiguous blocks of sequences having high Z-Scores and would be obvious when the Z-Score is plotted relative to the sequence number of the gene (as in Fig. 1). A Z-Score that takes the length of these longer sequences into account can be calculated by taking the average Z-Score of the extended block and correcting for the longer sequence by multiplying this average Z-Score by a factor N, where N is equal to the length of the contiguous block divided by the length of the search window. A standard window size for each analysis in this study, however, allows for the accurate comparison of Z-Scores from different genomic sequences.

Sample Z-Scores of some relatively simple sequences are shown in Table 1. These 6 to 8 dinucleotide sequences demonstrate the wide range of Z-Scores that are calculated by the Z-Hunt-II program. The table also shows the negative superhelicity required to induce the sequence to adopt the Z-conformation in a 5000-bp closed circular plasmid, as calculated by the Z-Hunt-II program. The Z-Scores in Table 1 differ from those the scores from the original Z-Hunt program (Ho *et al.*, 1986), because Z-Hunt-II more accurately incorporates the cooperativity of Z-DNA formation and quantifies the statistical occurrence of a given Z-DNA-forming sequence. Even though most of these sample sequences are APP, the Z-Scores vary over six orders of magnitude, and strongly indicate the bias toward C/G, and against A/T bps in Z-DNA. Based upon these sample scores, and their relationship to known sequences which adopt Z-DNA in supercoiled plasmids, we have set a minimum Z-Score of 1.0 for sequences which we consider to have a high probability for Z-DNA formation. This threshold also requires that all potentially "good" Z-DNA-forming sequences can adopt the left-handed conformation within a reasonable range of superhelical densities. For instance, in *Escherichia coli* it has been established that the superhelical density *in vivo* is between −0.025 and −0.04 (Rahmouni and Wells, 1989). The native superhelical density in eukaryotic cells, however, is yet to be determined.

Mapping Potential Z-DNA-forming Sequences in Human Genes—The Z-Hunt-II program was used to analyze human gene sequences retrieved from both the GenBank and EMBL sequence libraries. The DNA sequences were chosen for our study based upon the the following criteria: 1) they were complete human DNA sequences; 2) they contained an entire gene sequence coding for a known protein; and 3) they included some sequence from both the 5′- and 3′-flanking regions of the gene. We did not analyze cDNA sequences, pseudogenes, incomplete or partial sequences, or genes transcribed by either RNA polymerase I or III. Sequences for which insufficient data was available, either in the sequence file or the publication referenced within the file, to assign coding regions or other landmarks which were useful in categorizing the data were excluded from this analysis.

The plots shown in Fig. 1 are graphs of the Z-Hunt-II program analysis output for five different human genes. Each plot in Fig. 1 shows the Z-Scores *versus* the base pair number of each gene (numbered from the start of each individual sequence file). Additionally, on these plots we have diagrammed each gene to show the location of the primary transcript, and of the exons and introns in the gene. Potential Z-DNA-forming sequences in these genes correspond to peaks in the plots, and peaks of greater magnitude represent stronger Z-DNA-forming sequences (higher Z-Scores). The

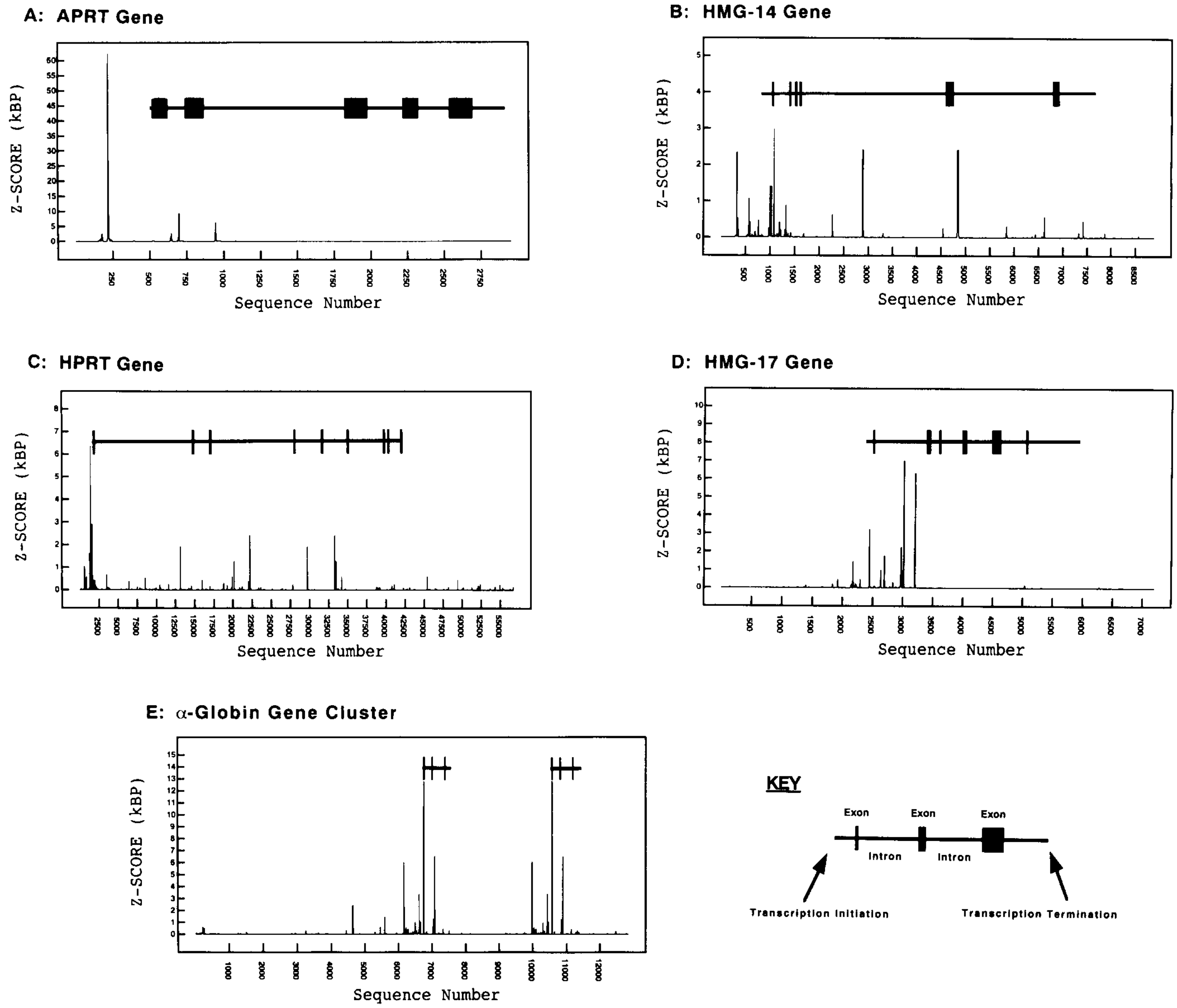

FIG. 1. **Distribution of potential Z-DNA-forming sequences in the APRT, HMG-14, HPRT, and HMG-17 genes and in the α-globin gene cluster.** Plots are of the Z-Score (in kilobase pairs) *versus* bp number for each gene. Note that the Z-Score values for each gene are individually scaled, therefore the scaling along the y-axes of these plots is not uniform. The *KEY* in the lower right, shows the symbols used in diagramming the approximate locations of the primary mRNA transcript, exons, and introns for each of the five genes shown.

first point that is apparent from the graphs in Fig. 1, is that potential Z-DNA-forming sequences are not limited to, nor are they excluded from, any region of the gene. Even in these few examples strong potential Z-DNA-forming sequences can be found in the 5′ regions of the genes and in both exons and introns. Although none of these particular genes have potential Z-DNA-forming sequences in their 3′-flanking regions, these types of sites were found in the analysis of other genes in the data set. We will discuss several other points concerning the distribution of Z-DNA in these five genes in later sections of the paper.

We have used Z-Hunt-II to search and map potential Z-DNA-forming sequences in a total of 137 complete human genes. In this data set, we found that 98 of the genes contained at least one strong, potential Z-DNA-forming sequence (Z-Scores ≥ 1.0). Therefore, 39 genes did not contain any sequences which had Z-Scores above this threshold level. The scope of the entire data set used in our analysis, as well as an overall summary of our results are shown in Table 2. The 1,003,901 bp of DNA sequence in the data set could be categorized as 5′ flanking to the transcription start site (TSS) of the gene (16%), introns (56%), and exons (14%) within the transcribed region of the gene, and 3′ flanking to the polyadenylation site of the gene (14%). In the entire data set we have identified and mapped 329 potential Z-DNA-forming sequences of 12–16 bp in length. This translates into an average of 1 potential Z-DNA-forming sequence every 3050 bps.

The complete listing of all the genes studied in this work is given in Tables 3 and 4. Table 3 lists the 39 genes which do not contain strong Z-DNA-forming sequences, as well as the length of each gene (in bp) and the locus identification number for the gene in either the Genbank or EMBL library. (The locus identification numbers for gene sequences retrieved from the GenBank data base begin with the prefix HUM-, while the gene sequences from the EMBL data base begin

TABLE 1

Z-scores of sample 6 to 8 dinucleotide sequences calculated with Z-Hunt-II

Also shown is the calculated superhelical density required to flip the sequence from B- to Z-DNA within a theoretical 5,000-bp plasmid (see "Materials and Methods").

Sequence	Required super-helical density	Z-Score
CGCGCGCGCGCGCGCGCG	−0.032	17,300
CGCGCGCGCGCGCGCG	−0.033	4,590
CGCGCGCGCGCGCG	−0.035	943
TGTGTGTGTGTGTGTGTG	−0.043	2.4
TGTGTGTGTGTGTGTG	−0.044	1.3
TGTGTGTGTGTGTG	−0.046	0.6
TATATATATATATATA	−0.061	0.003
TGCGTGCGTGCGTGCG	−0.038	135
TGCGTGCGTGCG	−0.040	16.7
CGTACGTACGTACGTA	−0.046	0.7
CGCCCGCGCCCG	−0.043	2.0
CGCCCGCGCCCGCCCG	−0.042	5.0

TABLE 2

Summary of data set and results

On average, one potential Z-DNA-forming sequence was found every 3,050 bps of total data set.

Total number of human genes studied	137
Genes containing Z-DNA	98
Genes without Z-DNA	39
Total length of all genes (bps)	1,003,901
Genes containing Z-DNA	856,259
Genes without Z-DNA	147,642
Percentage of data set in regions of genes	
5′-flanking regions	16%
Introns	56%
Exons	14%
3′-flanking regions	14%
Number of potential Z-DNA-forming sequences found in total data set with Z-Hunt-II	329

TABLE 3

List of the 39 human genes in this data set which do not contain potential Z-DNA-forming sequences (Z-Scores ≥ 1.0)

Gene	Locus identification	Length
		bp
Adenine nucleotide translocater-2	HSANT2X	4,982
α1-Antitrypsin	HSA1ATP	12,222
Atrial natriuretic factor	HSANF	2,710
Brain natriuretic protein	HSBNPA	1,922
Cathepsin G	HSCAPG	3,734
Cyclin	HSCYL	1,231
Cytokine LD78α	HSLD78A	3,176
Estradiol 17 β-dehydrogenase	HSEDHB17	4,845
Fatty acid binding protein	HSFABP	5,204
Granzyme B (CTLA-1)	HSCTLA1A	4,751
Granzyme H	HUMGHG	4,452
Heat shock protein 70	HSHSP70D	2,691
Histone H1°	HUMHIS10G	1,810
Histone H3	HSHIS3PR	1,125
Histone H2A	HUMHISH2A	866
Histone H2B	HUMHISH2B	843
Immunoglobulin germline κ chain V region	HSIGKVAA	1,331
Insulin	HUMINS01	4,044
α-Interferon	HUMIFNAD	1,179
Interleukin-1β	HSIL1B01	7,824
Interleukin-2	HUMIL2A	6,684
Interleukin-5	HSIL5	3,230
Islet amyloid polypeptide	HSHIAPPA	7,160
Keratin: type 1, epidermal	HUMKEREP	5,339
α-Lactalbumin	HUMLACTA	3,310
Matrix G1A protein	HSMGPA	7,734
Metallothionein I-F	HSMETIF1	2,076
Metallothionein-IG gene	HSMT2A	1,922
Monocyte chemotactic protein	HUMMCHEMP	2,776
Muscarinic acetylcholine receptor	HSCHRM	2,098
Phenylethanolamine *N*-methyl-transferase	HSPNMTA	4,174
Phosphoglycerate mutase	HSPGAMMG	3,771
Pulmonary surfactant apoprotein	HUMPSAP	4,778
Prealbumin	HSPALD	7,616
Protamine P1	HSPRT1A	304
Regenerating protein	HSREGA01	4,251
Serum amyloid A	HSSAA	3,460
Tumor necrosis factor	HUMTNFA	3,633
Tyrosinase	HUMTYRA	2,384

with the prefix HS-.) An alphabetical listing of the 98 human genes which do contain strong Z-DNA-forming sequences, as well as the length of the gene, and the Genbank or EMBL locus number are given in Table 4 in the Miniprint.[2] This extensive table also shows the six to eight dinucleotide sequences identified by the Z-Hunt-II program as having a high potential for Z-DNA formation, and the Z-Scores for these sequences, as well as the location of these sequences within the gene. In Table 4 we have defined the location of the Z-DNA-forming sequences as both: 1) the region of the gene where the sequence is located (*i.e.* in an intron or promoter region, etc.) and 2) the percentile of the site from the beginning of the sequence file (*i.e.* if a 1000-bp sequence file has a Z-DNA-forming sequence at the 21.5 percentile, then the sequence is located 215 bp from the start of the sequence file).

Many of the Z-DNA-forming sequences identified by the Z-Hunt-II program are poly(GT/AC)$_n$ type sequences (see Table 4). These sequences, where n = 10–60, are thought to be repeated up to 50,000 times in the human genome (Hamada *et al.*, 1982a, 1982b; Gross and Garrard, 1986), and have been shown to form Z-DNA in negatively supercoiled plasmids *in vitro* (Nordheim and Rich, 1983; Hanniford and Pulleyblank, 1983; Hayes and Dixon, 1985; Naylor and Clark, 1990). However, Z-Hunt-II also flags many other, perhaps less obvious potential Z-DNA-forming sequences. Z-Hunt-II is especially useful for identifying sequences which are not 100% APP and would therefore be overlooked using an algorithm which searches exclusively for APP sequences. This is an important feature, since several other studies have in the past equated all APP sequences with potential Z-DNA formation, even those which are prohibitively A/T-rich. The unique aspect of a thermodynamic search strategy as opposed to previous sequence matching algorithms is that it ranks sequences according to their structural propensity to form Z-DNA, regardless of whether they are strictly APP sequences. For example, the sequence CGCCCGCGCCCGCCCG, which has 3 cytosine residues (*underlined*) which are out of alternation, is assigned a Z-Score of 5.0 (see Table 1). The energetic cost of the 3 out-of-alternation residues in this sequence is included in the calculation performed by Z-Hunt-II, whereas this sequence would be overlooked if one were using a search strategy which identifies only APP sequences.

The extensive data in Table 4 are intended to provide a good starting point for comparison with other analyses of

[2] Part of this paper (Table 4) is presented in miniprint at the end of this paper. Miniprint is easily read with the aid of a standard magnifying glass. Full size photocopies are included in the microfilm edition of the Journal that is available from Waverly Press.

DNA sequences. These data allow one to put other potential Z-DNA-forming sequences, perhaps those located in genes which have yet to be sequenced, into the context of this rather large data set. It will also eventually be important to compare the location of potential Z-DNA-forming sequences in these same genes studied across phylogenetic lines, to determine if these structures are evolutionarily conserved. In addition, we are hopeful that the results in Table 4 will interest workers studying transcriptional regulation of some of these specific genes in which Z-DNA could possibly play an important regulatory role. For instance, the cytoplasmic β-actin gene contains several very strong Z-DNA-forming sequences in its 5′ UTR (see Table 4), which in this gene is highly conserved between humans and bovine (Nakajima-Iijima *et al.*, 1985). It is possible that genes such as this may be good candidates for studying the effects of Z-DNA on eukaryotic gene regulation, an area which is poorly understood.

The 20 sequences found to have the highest thermodynamic probability for Z-DNA formation in this data set of human DNA are given in Table 5. The sequence which received the highest Z-Score (TGCGTGCGCGCGCGCG, Z-Score = 1750), was located in the 5′-untranslated region of the cytoplasmic β-actin gene. The two highest possible scoring sequences for a 6–8-bp window size, $(GC)_7$ and $(GC)_8$ as shown in Table 1, were not found in any of the genes studied. This is consistent with other results showing that the sequences GCGCGCGC and CGCGCGCG are significantly underrepresented in eukaryotic genomes (Trifonov *et al.*, 1985). The other top Z-DNA-forming sequences found in human genes are, in general, APP combinations of GC and GT or AC dinucleotides. We should note, however, that outside this narrow set of sequences, a large number of non-APP sequences were identified as having strong propensities for adopting the Z conformation.

Distribution of Z-DNA-forming Sequences across Human Genes—In an effort to analyze the distribution of potential Z-DNA-forming sequences across human genes, we have tabulated the location of all the Z-DNA-forming sequences listed in Table 4. These results are shown in Table 6 and Fig. 2. The locations of the Z-DNA-forming sequences have been categorized in Table 6 according to their location within well defined regions of the gene. For this analysis we have divided the genes into the following categories: 5′-flanking region, promoter region, 5′ UTR, exons, introns, 3′ UTR, and 3′-flanking region. We find that ~35% of the Z-DNA sites are located upstream, or 5′, to the first expressed exon in these genes (*i.e.* in either the 5′-flanking region, promoter region, or 5′ UTR). Conversely, only 3% of the Z-DNA sites are found downstream, or 3′, to the last expressed exon in these genes (*i.e.* in either the 3′ UTR or the 3′-flanking region). These regions represented nearly identical percentages of the sequenced DNA in the data set (16% in the 5′-flanking regions, and 14% in the 3′-flanking region). Thus, the number of strong Z-DNA-forming sequences occur nearly twice as frequently in the upstream regions as one would suspect from the number of base pairs in the region. Potential Z-DNA-forming sequences are particularly enriched in the 5′-untranslated regions of these genes, which contained 8.8% of these sequences (see Table 6), yet comprise probably less than 1% of the total sequence in the data set. The remaining 62% of the potential Z-DNA sites were located either in exons (14.9%) or introns (47.1%). This is in general accord with expectations from the percentage of base pairs represented by these regions (exons accounting for 14% and introns accounting for 56% of the data set).

The nonrandom distribution of potential Z-DNA-forming sequences across the human genes in our data set is further

TABLE 5
Top 20 thermodynamically most stable Z-DNA-forming sequences found in the entire data set of 137 complete human genes using the Z-Hunt-II program

Sequence	Z-Score	Gene
1. TGCGTGCGCGCGCGCG	1750	β-Actin, cytoplasmic
2. GCGCCCGCGCGCGCGC	733	Factor VII
3. GCGCGCGCGCGT	303	Desmin
4. GCGCGTGCGCGC	199	β-Actin, cytoplasmic
5. CGCGCGCGCGCCCATG	148	Apolipoprotein A-I
6. GCACGCACACGCGCGT	132	*int*-1 (c-*myc*)
7. GCGCGCGCGCGG	129	L-*myc*
8. CGCACGCGCACGCA	103	*int*-1 (c-*myc*)
9. CGCGCGCGCACA	75.0	α-Actin, skeletal
10. TGTGTGCGCGCGTGTG	71.2	β-Globin gene region
11. TGTGCGCGCGCACATG	71.2	Pulmonary surfactant C
12. GCGCGCCCGTACGCGC	59.0	APRT
13. GCGCACGCACGC	48.8	c-*fos*
14. CGCGCACGCACACATG	46.4	Erythropoietin
15. GCGCGCACGCGGACAC	39.4	*hst* protein
16. CGCGCGCGCCCG	37.9	Ubiquitin-like protein
17. GCGCGCGCCCGC	37.9	*int*-2
18. CACGCGCACGTGCCCG	37.1	Cytochrome P450-IID6
19. GTGCGTGCCCGCGCGT	35.0	Adenylate kinase
20. CGTGCGTGTGTGTGCG	31.7	Adenylate kinase

TABLE 6
Distribution of potential Z-DNA-forming sequences in 137 human genes

Location	Sequences	
	No.	% total
5′-Flanking regions	22	6.7
Promoter regions	62	18.8
5′-Untranslated region	31	9.4
Introns	155	47.1
Exons	49	14.9
3′-Untranslated regions	2	0.6
3′-Flanking regions	8	2.4
Total	329	

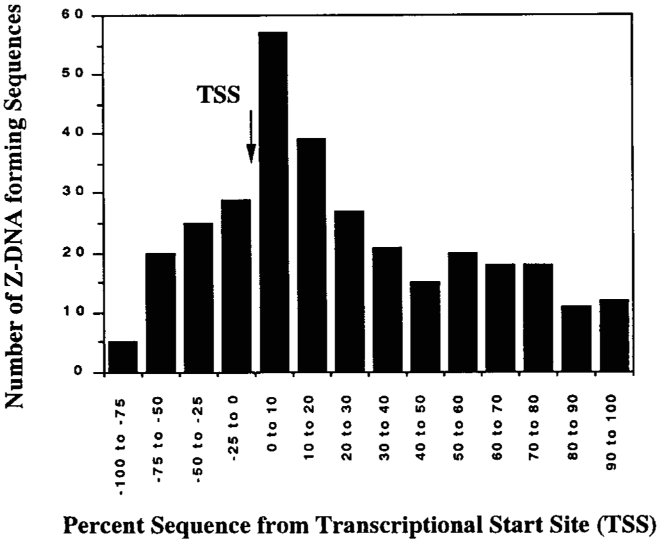

FIG. 2. **Distribution of potential Z-DNA-forming sequences across 137 human gene sequences.** A graph of the number of potential Z-DNA-forming sequences *versus* the percentage of base pairs in each gene that are upstream (5′) and downstream (3′) to the TSS of the gene. The potential Z-DNA-forming sequences have been grouped into larger percentages in the 5′ end to illustrate that this region represents a smaller proportion (16%) of the total data set, relative to the amount of sequence in the data set which is 3′ to the TSS (84%).

emphasized in the graph shown in Fig. 2. In this graph, the potential Z-DNA-forming sequences have been plotted relative to the transcription start site for each of the genes. The base sequences are shown as the percentage of base pairs in the data set that are upstream and downstream of the TSS of each gene. This makes the analysis of the entire data set relative to a common functional feature possible, even though each gene in the data set is of different length. The percentages on this axis are labeled in a manner that mirrors the number of bases that are upstream (16%) *versus* downstream (84%) of the TSS. When plotted this way, it is clear that potential Z-DNA-forming sequences are not randomly distributed across the human genes of the data set. Of the 317 potential Z-DNA-forming sequences used in generating Fig. 2, 79 of these (representing 25% of the total) were located upstream from the TSS, and 94 (30% of total) were found within 20% of the sequence downstream of the TSS. Together these two categories (5′-flanking plus first 20% downstream of TSS) represent over 54% of the total number of strong Z-forming sequences. It is also clear from Fig. 2 that downstream of the TSS in these genes, the potential Z-DNA-forming sequences are strongly favored at the 5′ end of the gene (159 sequences from 0 to 50% relative to the TSS) over the 3′ end (79 from 50 to 100%).

The overall nonrandom distribution of potential Z-DNA-forming sequences in human genes shown in Table 6 and Fig. 2, is also evident in the analysis of the individual genes shown in Fig. 1. The five plots in Fig. 1 show the Z-Hunt-II analysis of the APRT, HMG-14, HPRT, HMG-17 genes and the α-globin gene cluster. The distribution of Z-DNA-forming regions in these five gene sequences are excellent examples of the distribution of potential Z-DNA-forming sequences in the human genes studied in this data set. In these cases, most of the potential Z-DNA-forming sequences (Z-Scores ≥ 1.0) are located either in the promoter region or close to the site of transcription initiation. The clustering of Z-DNA-forming sequences near the start of transcription is clearly evident in the four genes shown in Fig. 1, *A-D*. Furthermore, even the Z-DNA-forming sequences which had Z-Scores below our cutoff of 1.0 seem to be more enriched in the 5′ regions of these genes.

The human DNA sequence analyzed in *panel* E of Fig. 1 contains two α-globin genes, and gives results which hint of another potentially intriguing feature concerning the distribution of Z-DNA-forming sequences in the human genome. Since this sequence file contains more than one gene, it is one of the few files which contains true "intergenic" DNA sequences. Note that the region of DNA between the two transcribed genes is particularly devoid of any Z-DNA-forming sequences. This phenomenon was also observed in the 73,326-bp β-globin gene cluster, which contains five different genes (see Table 4), and in the analysis of the HMG-17 gene, which contains a higher proportion of 5′-flanking sequences than most of the genes in our study (Fig. 1*D*), as well as in several other instances. At this point, because relatively small amounts of noncoding DNA sequence information is available, we can only suggest that Z-DNA-forming sequences may be relatively depleted in intergenic DNA sequences. It is interesting, that in Fig. 1*E* the Z-DNA-forming sequences align very well with the location of the two α-globin genes, and that the distribution of the Z-DNA-forming sequences is centered at the site of transcription initiation. Perhaps also interesting, is that an α-globin pseudogene is located between 2,436 and 3,248 bp in this sequence, which is a region without the Z-DNA-forming sequences found in and around the two functional α-globin genes (see Fig. 1*E*). Further analysis of larger genomic DNA sequences, including noncoding and intergenic sequences, will be required to develop these ideas and to formulate any rules which may apply to the global distribution of potential Z-DNA-forming sequences in the human genome.

DISCUSSION

A study such as this one may be a useful approach to begin addressing questions of what, if any, role does DNA structure and structural transitions play in the biochemistry of the cell. Because the physical chemistry of the transition between right-handed B-DNA and left-handed Z-DNA is relatively well understood (Jovin *et al.*, 1987), we are able to use a computer program (Z-Hunt-II) to predict and map the occurrence of potential Z-DNA-forming regions in genomic sequences, based upon their thermodynamic propensity to form Z-DNA. Our results show that potential Z-DNA-forming sequences are found in and around many human genes (Table 4), and that these sequences are nonrandomly distributed, showing a strong tendency to be located close to the site of transcription initiation in the gene (Table 6 and Fig. 2). Several previous studies have noted the presence of potential Z-DNA-forming sequences in both prokaryotic and eukaryotic genomes (Konopka *et al.*, 1985; Braaten *et al.*, 1985; Hoheisel and Pohl, 1987; Rollo *et al.*, 1989), but this is the first study which maps the occurrence of potential Z-DNA-forming sequences in such a large, self consistent, data set.

Until recently a major question concerning Z-DNA formation in eukaryotic cells has always been, do eukaryotic cells have enough negative supercoiling to facilitate Z-DNA formation *in vivo*? It is well established that negative supercoiling is an energetic prerequisite for Z-DNA formation in physiological conditions. However, it was shown only in the last few years that negative supercoiling could be generated by normal DNA processing events, such as transcription, and did not require the enzymatic activity of a eukaryotic gyrase (which has yet to be found). Liu and Wang's elegant twin-supercoiled-domain model proposed that actively transcribing RNA polymerase complexes will generate positive supercoiling in front of, and negative supercoiling behind, the elongating polymerase (Liu and Wang, 1987). Subsequent work by them and others has shown that transcription does indeed generate relatively high levels of dynamic negative and positive supercoiling (Wu *et al.*, 1988; Brill and Sternglanz, 1988; Figueroa and Bossi, 1988; Giaever and Wang, 1989; Tsao *et al.*, 1989). In addition, the transcription-induced supercoiling phenomenon seems to be quite general, as it has been observed both *in vitro* and *in vivo*, and in prokaryotes and eukaryotes. Liu and Wang (1987) suggested in their original paper describing the twin-domain model of transcription-induced supercoiling, that this process could potentially generate enough negative supercoiling energy to drive the formation of non-B-DNA structures near the promoters of actively transcribing genes. This idea has been subsequently supported by results showing that Z-DNA-forming sequences placed upstream of an actively transcribing gene will form Z-DNA, but those located downstream of the gene will not flip to Z-DNA (Rahmouni and Wells, 1989; Droge and Nordheim, 1991).

The location of many of the sequences flagged by our analysis as having high potential for Z-DNA formation correlates with regions of the gene which are expected to be, at least transiently, negatively supercoiled during transcription. This is a very interesting correlation since, as discussed before, negative supercoiling is required to facilitate the local flipping of regions of B-DNA to Z-DNA in constrained DNA molecules. Our data mapping potentially strong Z-DNA-forming sequences near genes also strongly correlate with other work showing that anti-Z-DNA antibodies selectively bind to actively transcribing regions of chromosomes (Lancilloti *et al.*, 1987; Jimenez-Ruiz *et al.*, 1991), and that the binding of anti-Z-DNA antibodies to permeabilized mammalian nuclei is largely dependent upon transcriptional activity (Wittig *et al.*, 1989, 1991). Our current view of the transcriptional process as having a very dynamic effect on local DNA topology, has led to the suggestion that transcription is one of the principle factors affecting DNA supercoiling in eukaryotic cells (Liu and Wang, 1987; Wu *et al.*, 1988; Wittig *et al.*, 1991). Because of this, it is also possible that transcription provides much of the negative supercoiling energy required for Z-DNA formation in mammalian cells (Wittig *et al.*, 1991).

Although the flux of negative supercoiling density in the wake of a transcription complex has been shown to have both biological and structural effects, including facilitating Z-DNA formation (Rahmouni and Wells, 1989; Droge and Nordheim, 1991), it is difficult to predict which of the potential Z-DNA-forming sequences mapped with Z-Hunt-II would be in a highly negative supercoiled environment during transcription because of the extremely dynamic nature of the transcription process. The 329 potential Z-DNA-forming sequences which we have identified (listed in Table 4), are simply the most thermodynamically favorable Z-DNA-forming sequences in this data set, as identified with the Z-Hunt-II program. The local superhelical density at any given site along the DNA will ultimately determine the ability of these sequences to adopt the Z-DNA conformation. Local negative supercoiling depends upon many complex factors (Liu and Wang, 1987) including the rate of transcription, the number of active transcription complexes, local topoisomerase activity, the binding of specific proteins, and the chromatin structure of the region (discussed further by Liu and Wang, 1987; Wu *et al.*, 1988; Giaever and Wang, 1989; Pfaffle *et al.*, 1990; Lee and Garrard, 1991). Also it is important to consider that the DNA in eukaryotic chromosomes is organized into linear DNA loop domains (Cockerill and Garrard, 1986; Gross and Garrard, 1987), and therefore it is these chromosomal structures which define the local topological domain of any human DNA sequence. Chromosomal loop domains vary in size from about 5 to 100 kilobase pairs, are thought to contain a single gene or a gene cluster, and are tethered on each side by very strong interactions with the nuclear matrix (Cockerill and Garrard, 1986; Gross and Garrard, 1987).

The effect of twin-domain, transcription-induced supercoiling on the topology of a eukaryotic chromosomal loop domain has not yet been investigated. All of the past experimental validations of the twin-domain model have studied transcription-induced supercoiling in small, closed circular DNA molecules (plasmids), in which the entire DNA molecule is topologically linked (Wu *et al.*, 1988; Brill and Sternglanz, 1988; Figueroa and Bossi, 1988; Giaever and Wang, 1989; Tsao *et al.*, 1989). In the case of a single gene being transcribed on a circular template, the negative supercoils behind and the positive supercoils in front of the polymerase can migrate around the circular template, and eventually cancel each other. But, in the case of a single gene located in a linear chromosomal loop domain anchored on each end by strong interactions with the nuclear matrix, the positive and negative supercoils generated by transcription cannot easily migrate toward one another to cancel each other, until transcription ceases. Because of this, one could imagine that the negative supercoiling density at the 5′ end of a eukaryotic gene, contained within a chromosomal loop domain, could potentially reach very high levels. This type of linearly constrained DNA organization increases the combination of factors which could influence transcription-induced local superhelical fluctuations, when compared to the relatively well studied effects of transcription on circular DNA templates. For instance, the distance (in bp) between the transcription start site of a gene and the nuclear matrix attachment site could be an important factor in determining the extent of negative supercoiling in some promoter regions.

In summary, we have mapped the 329 most thermodynamically favorable Z-DNA-forming sequences in 137 different human genes. We find that the distribution of these sequences relative to the location of the genes is distinctly nonrandom, being strongly enriched in regions of the gene which are expected to be transiently negatively supercoiled during transcription. This interesting correlation suggests a mechanism whereby many of these sequences could flip to Z-DNA resulting from transcription of the gene in which they are located. Determining the possible function of flipping small regions of B-DNA to Z-DNA in and around transcribed regions of the human genome is undoubtedly going to be a challenging problem, since the B-Z transition in these sequences is likely dependent upon dynamic transcription-induced supercoiling. However, the nonrandom nature of the distribution of Z-DNA across human genes is suggestive of a possible role for Z-DNA in transcriptional regulation.

REFERENCES

Braaten, D. C., Thomas, J. R., Little, R. D., Dickson, K. R., Goldberg, I., Schlessinger, D., Ciccodicola, A., and D'Urso, M. (1988) *Nucleic Acids Res.* **16,** 865–881

Brill, S. J., and Sternglanz, R. (1988) *Cell* **54,** 403–411

Cockerill, P. N., and Garrard, W. T. (1986) *Cell* **44,** 273–282

Droge, P., and Nordheim, A. (1991) *Nucleic Acids Res.* **19,** 2941–2946

Ellison, M. J., Kelleher, R. J., III, Wang, A. H.-J., Habener, J. F., and Rich, A. (1985) *Proc. Natl. Acad. Sci. U. S. A.* **82,** 8320–8324

Figueroa, N., and Bossi, L. (1988) *Proc. Natl. Acad. Sci. U. S. A.* **85,** 9416–9420

Giaever, G. N., and Wang, J. C. (1988) *Cell* **55,** 849–856

Gilbert, W. (1991) *Nature* **349,** 99

Gross, D. S., and Garrard, W. T. (1986) *Mol. Cell. Biol.* **6,** 3010–3013

Hamada, H., and Kakunaga, T. (1982a) *Nature* **298,** 396–398

Hamada, H., Petrino, M. G., and Kakunaga, T. (1982b) *Proc. Natl. Acad. Sci. U. S. A.* **79,** 6465–6469

Haniford, D. B, and Pulleyblank, D. E. (1983) *Nature* **302,** 632–635

Hayes, T. E., and Dixon, J. E. (1985) *J. Biol. Chem.* **260,** 8145–8156

Ho, P. S., Ellison, M. J., Quigley, G. J., and Rich, A. (1986) *EMBO J.* **5,** 2737–2744

Hoheisel, J. D., and Pohl, F. M. (1987) *J. Mol. Biol.* **193,** 447–464

Jaworski, A., Hsieh, W.-T., Blaho, J. A., Larson, J. E., and Wells, R. D. (1987) *Science* **238,** 773–777

Jimenez-Ruiz, A., Requena, J. M., Lopez, M. C., and Alonso, C. (1991) *Proc. Natl. Acad. Sci. U. S. A.* **88,** 31–35

Jovin, T. M., Soumpasis, D. M., and McIntosh, L. P. (1987) *Annu. Rev. Phys. Chem.* **38,** 521–560

Kagawa, T. F., Stoddard, D., Zhou, G., and Ho, P. S. (1989) *Biochemistry* **28,** 6642–6651

Kennard, O., and Hunter, W. N. (1989) *Q. Rev. Biophys.* **22,** 327–379

Konopka, A. K., Reiter, J., Jung, M., Zarling, D. A., and Jovin, T. M. (1985) *Nucleic Acids Res.* **13,** 1683–1701

Lancillotti, F., Lopez, M. C., Arias, P., and Alonso, C. (1987) *Proc. Natl. Acad. Sci. U. S. A.* **84,** 1560–1564

Lee, M.-S., and Garrard, W. T. (1991) *Proc. Natl. Acad. Sci. U. S. A.* **88,** 9675–9679

Liu, L. F., and Wang, J. C. (1987) *Proc. Natl. Acad. Sci U. S. A.* **84,** 7024–7027

Maddox, J. (1991) *Nature* **352,** 11–14

McClellan, J. A., Palecek, E., and Lilley, D. M. J. (1986) *Nucleic Acids Res.* **14,** 9291–9309

McLean, M. J., Lee, J. W., and Wells, R. D. (1988) *J. Biol. Chem.* **263,** 7378–7385

McLean, M. J., and Wells, R. D. (1988) *Biochim. Biophys. Acta* **950,** 243–254

Nakajima-Iijima, S., Hamada, H., Reddy, P., and Kakunaga, T. (1985) *Proc. Natl. Acad. Sci. U. S. A.* **82,** 6133–6137

Naylor, L. H., and Clark, E. M. (1990) *Nucleic Acids Res.* **18,** 1595–1601

Nordheim, A., and Rich, A. (1983a) *Nature* **303,** 674–679

Nordheim, A., and Rich, A. (1983b) *Proc. Natl. Acad. U. S. A.* **80,** 1821–1825

Panyutin, I., Lyamichev, V., and Mirkin, S. (1985) *J. Biomol. Struct. & Dyn.* **2,** 1221–1232

Peck, L. J., and Wang, J. C. (1983) *Proc. Natl. Acad. Sci. U. S. A.* **80,** 6206–6210

Pfaffle, P., Gerlach, V., Bunzel, L., and Jackson, V. (1990) *J. Biol. Chem.* **265,** 16830–16840

Rahmouni, A. R., and Wells, R. D. (1989) *Science* **246,** 358–363

Rollo, F., Amici, A., and Mancini, G., (1989) *J. Mol. Evol.* **28,** 225–231

Rich, A., Nordheim, A., and Wang, A. H.-J. (1984) *Annu. Rev. Biochem.* **53,** 791–846

Trifonov, E. N., Konopka, A. K., and Jovin, T. M. (1985) *FEBS Lett.* **185,** 197–202

Tsao, Y.-P., Wu, H.-Y., and Liu, L. F. (1989) *Cell* **56,** 111–118

Wang, J. C., and Giaever, G. N. (1988) *Science* **240,** 300–304

Wang, A. H.-J., Quigley, G. J., Kolpak, F. J., Crawford, J., L., van Der Marel, G. A., van Boom, J. H., and Rich, A. (1979) *Nature* **282,** 680–686

Wells, R. D. (1988) *J. Biol. Chem.* **263,** 1095–1098

Wittig, B., Dorbic, T., and Rich, A. (1989) *J. Cell Biol.* **108,** 755–764

Wittig, B., Dorbic, T., and Rich, A. (1991) *Proc. Natl. Acad. Sci. U. S. A.* **88,** 2259–2263

Wu, H.-Y., Shyy, S., Wang, J. C., and Liu, L. F. (1988) *Cell* **53,** 433–440

SUPPLEMENTARY MATERIAL

MAPPING Z-DNA IN THE HUMAN GENOME:

Computer Aided Mapping Reveals a Non-Random Distribution of Potential Z-DNA Forming Sequences in Human Genes

By Gary P. Schroth, Ping-Jung Chou, and P. Shing Ho

Table 4

List of the 98 human genes which contain potential Z-DNA forming sequences. The table also shows the 6 to 8 dinucleotide potential Z-DNA forming sequences identified by the Z-Hunt-II program, the Z-Scores for each sequence, and location of the sequences within each gene.

Name	Length	Locus	Sequence	Z-Score	Location	(%)
α-1-Acid Glyco-Protein (AGP2)	4,944	HSA1GLY2	CCTGCATGCGCA	1.0	Intron	(40.6)
α-Actin, Skeletal	3,778	HUMSAACT	CGCGCGCGCACA	75.0	Intron	(62.3)
			GCGCAGGCGCACACATGC	5.0	Promoter	(9.9)
			TGATGCACGCGCCTCT	2.1	Intron	(43.2)
			CACAGGTGCGCG	1.0	Intron	(66.8)
β-Actin, Cytoplasmic	3,646	HSACCYBB	TGCGTGCGCGCGCGCG	1750	5' UTR	(22.2)
			GCGCGTGCGCGC	199	5' UTR	(17.7)
			GCGCGTGCGCGC	199	5' UTR	(19.3)
			CTGCGTGCGCGC	21.6	5' UTR	(18.5)
			CGCGCGTGCGTT	20.0	5' UTR	(20.2)
			CTGGGCGCGCGC	9.3	5' UTR	(27.0)
			GCGGGCGCGCTC	3.5	5' UTR	(13.8)
			GCCCGCGAGCAC	2.6	5' UTR	(6.9)
			CGCGCTCGGGCGGGCG	1.2	5' UTR	(13.6)
			ATCAGCGTGCGC	1.2	5' UTR	(4.3)
γ-Actin, Cytoskeletal	3,583	HUMACTGA	CCGCGCGCGCCG	10.0	Promoter	(11.3)
			CCGCGGGCGCGC	4.6	Promoter	(11.8)
			CGAGCGCGCGGG	1.3	Promoter	(7.2)
			CAGACGCGCCCGCCTG	1.1	Intron	(24.6)
APRT: Adenine Phosphoribosyltransferase	2,956	HUMAPRTA	GCGCGCCCGTACGCGC	59.0	Promoter	(7.4)
			GCGCGTGCCCGC	9.5	Intron	(23.6)
			CTCTGCGCGCGC	6.3	Intron	(31.9)
			CCCGCGCGCAGC	2.5	Promoter	(6.0)
			GTGCACGCACAG	2.0	Intron	(21.9)
Adenosine Deaminase	36,741	HSADAG	GCGCGCGCCCAC	17.3	Promoter	(10.1)
			GCGCGCGCGGGGGC	15.4	Intron	(11.1)
			TGTGTGCATGTGTGCA	2.4	Intron	(70.9)
			CACACACACACACACA	2.4	Intron	(71.0)
			GCAGGCACACAGGC	1.6	Intron	(46.0)
			GTGCGGGCGCCT	1.1	Intron	(15.7)
Adenylate Kinase	12,229	HUMAK1	GTGCGTGCCCGCGCGT	35.0	Intron	(8.1)
			CGTGCGTGTGTGTGCG	31.7	Intron	(9.7)
			CACACACGCACACACG	15.1	Intron	(60.7)
			CGCGTGTGTGTGGGCA	5.4	Promoter	(5.8)
			TGCACACGTGCTCGCA	1.9	Intron	(61.9)
			GCATGTGCACAC	1.0	Intron	(61.3)
α-2-Adrenergic Receptor	3,604	HSADRA2R	GGGCGCGGGCGC	3.2	5' Flank	(25.8)
			CGCACCCGCGTG	3.0	Exon	(76.4)
			GCGGGCGCCCGCGT	1.8	5' Flank	(55.9)
			CGTGCACCTGTGCG	1.6	Exon	(67.7)
			CGCGCTGGCGCGGGCG	1.3	Exon	(87.5)
			GCGCCCGCGTAG	1.1	3' Flank	(95.5)
			CGCGTGCCGCCG	1.0	5' Flank	(30.7)
β-2-Adrenergic Receptor	3,458	HSADRBRA	GCGCGCGCGAGTGTGC	17.1	Promoter	(15.5)
			GCGCAGGCGCCC	1.8	Promoter	(21.0)
			GTATGTGCGTGC	1.1	Promoter	(18.5)
Serum Albumin	19,002	HSALBGC	GCATGCACGTGTGTGT	6.8	Intron	(80.2)
			GTGTGTGTGCATGCGT	6.8	Intron	(6.5)
			TGCACACACACACACA	2.4	Intron	(34.7)
Alkaline Phosphatase	4,556	HSALPHA	TGTTCGCGCGCG	5.9	Exon	(70.0)
			CACACACACACACACA	2.4	Intron	(11.9)
α-fetoprotein	22,166	HSAFPCP	CAGGCACGCGCC	1.9	3' Flank	(99.7)
			TGTGTGTGTGTGTG	1.3	Intron	(20.6)
Angiogenin	4,668	HSAGG	TGCGCATGTGCG	4.6	Promoter	(17.3)
			CACACACACACACACA	2.4	Intron	(66.6)
			ACAGGCGCCCGCAC	1.1	Promoter	(26.5)
ALP* A-I	2,385	HUMAPOAIT	CGCGCGCGCGCCCATG	148	Exon	(73.2)
			GAGGGCGCGCGC	9.2	Exon	(70.3)
ALP* A-II	2,928	HUMAPOA2I	TGTGTGTGTGTGTGTG	2.4	Intron	(49.4)
ALP* A-IV	3,613	HUMAPOA4A	GCACGTGTGCAT	1.8	Intron	(61.5)
			GTGCCCACGCACGG	1.2	Promoter	(18.6)
ALP* C-I	5,375	HSAPOCIA	GCACGCGCCTGT	1.9	Intron	(30.3)
			GCGCCCGCCTCCGCGC	1.3	Intron	(23.2)

Table 4, Cont.

Name	Length	Locus	Sequence	Z-Score	Location	(%)
ALP* C-II	4,057	HSAPOC2B	GTGTGTGTGTGTGTGT	2.4	Intron	(16.4)
			TGTGTGTGTGTGTGTG	2.4	Intron	(17.1)
ALP* E	5,515	HUMAPOE4	CTGCGCGCGCGG	10.0	Exon	(76.9)
			CGCGTGCGGGCC	2.5	Exon	(75.5)
			GCGGAGGTGCGCGC	1.9	Exon	(78.1)
			ACGTGCGCGGCCGC	1.8	Exon	(71.1)
Na,K-ATPase, α-2 Subunit	26,668	HUMATP1A2	CGCATGCGCACACACA	15.2	Intron	(42.6)
			CACACACACGCACACA	6.8	Intron	(71.3)
			ACGCACACACACATGC	3.6	Intron	(72.2)
			CACACACACACACACA	2.4	Intron	(71.9)
			ACACACACACACACAC	2.4	Intron	(71.5)
			AGCGCACATGTGTG	1.4	Intron	(81.2)
			ACACACACACACAC	1.3	Intron	(71.4)
ATP Synthase, β-Subunit	10,186	HSATPSYB	GTGTGTGTGTGTGTGT	2.4	5' Flank	(3.4)
			ACAGGTGCGCGC	2.2	Intron	(82.8)
			GTGCGGGCAGGTGCGT	1.0	Intron	(24.2)
c-*abl* Oncogene	3,840	HSABLA	GCGGGCGCGGGC	2.0	Promoter	(5.7)
			GTGCGGACGGGCGC	1.3	Promoter	(2.6)
Calcyclin	3,671	HSCACY	CCCACGTGCATGCACAT	3.6	Promoter	(21.1)
c-*fos* Oncogene	6,210	HUMFOS	GCGCACGCACGC	48.8	Intron	(19.2)
			CACACACATGCACACG	5.2	3' UTR	(85.9)
			AGGCCCGCGCGC	3.5	Promoter	(10.1)
c-*syn* Oncogene	2,647	HSCSYNA	GCCCGGGCGCACAC	2.9	5' UTR	(8.4)
			CGCCCGGGCGCC	2.1	5' UTR	(6.9)
C-Reactive Protein	2,438	HUMCRPG	GTGTGTGTGTGTGTGT	2.4	Intron	(18.8)
Cys-Proteinase Inhibitor (CST1)	3,716	HSINCP	GTGTGTGTGTGTGC	3.0	Intron	(61.3)
			ACACGTGTGTACACAC	2.4	Intron	(64.1)
			GTGCACGCAGGC	1.1	3' Flank	(98.9)
Cytochrome C, Somatic	3,088	HUMCYCAA	GCGCGCACTTGC	1.8	Intron	(6.3)
Cytochrome C-1	4,622	HUMCYC1A	CGGGCGCGCGTGCCCG	23.6	Exon	(31.8)
			ACAGGCGCGCGCCACC	8.0	5' Flank	(11.4)
			GTGCACGCGCTG	1.2	5' Flank	(20.7)
Cytochrome P450-C21	5,141	HSMHCP42	GCACGTGCACAT	2.1	Exon	(65.3)
Cytochrome P450-IID6	9,432	HSCYP2D6	CACGCGCACGTGCCCG	37.1	Intron	(33.9)
			CGGCCGTGCGCG	1.7	Exon	(27.3)
Cytochrome P450-IIE1	14,776	HSCYPIIE	TTGCGCGTGCGC	19.5	Intron	(23.6)
			GTGCGCACGTGC	16.8	Intron	(58.3)
			CGCGCGCGGGCC	10.5	Intron	(26.2)
			CGTGCGTGCGGCTGCA	4.9	Intron	(63.1)
			GAGCGTGCGCTC	1.5	Intron	(34.7)
			GCACGCGCCTAT	1.5	Intron	(71.8)
			GTGGGCGCGCCT	1.4	Intron	(24.4)
CK8: Cytokeratin 8	8,815	HSDKERB	TGTGTGTGCGCGTGTG	20.9	Promoter	(4.2)
			GTGCGTGCACAG	2.0	Intron	(36.0)
			CGTGCGCACCCA	1.3	Exon	(15.5)
			ACAGGCGCCCGCAC	1.1	Intron	(77.5)
Desmin	8,878	HSDES	GCGCGCGCGCGT	303	Exon	(7.6)
			GCACACACATGCACAC	3.1	Intron	(45.1)
			GCGTGCGCCGCGCCCG	2.0	Intron	(41.1)
			CGCGCACGTCGG	1.5	Exon	(3.8)
			CCGTGCGCCCGCCAGC	1.5	5' UTR	(1.5)
EF-1A: Elongation Factor 1-α	4,695	HUMEF1A	CGCGTGCGAATCTG	1.2	Intron	(17.5)
Eosinophil Major Basic Protein	3,608	HSEMBPA	TGTGTGTGCATGTGTG	2.4	Intron	(25.8)
Erythropoietin	3,602	HSERPA	CGCGCACGCACACATG	46.4	Promoter	(9.4)
			CACACGCACGTCTGCA	2.0	Promoter	(5.7)

Table 4, Cont.

Name	Length	Locus	Sequence	Z-Score	Location	(%)
Factor VII	12,850	HSCFVII	GCGCCCGCGCGCGCGC	733	Intron	(35.6)
			CGCAGACGCGCGCG	15.1	Intron	(37.3)
			GTGCGCACACACAC	11.6	Intron	(96.3)
			GCACACACACACACGC	10.9	Intron	(94.3)
			TGCACACACACACG	6.0	Intron	(95.9)
			TGTGCGCACACACA	6.0	Intron	(96.0)
			CGCGCGTGCCCCCG	4.5	Intron	(31.4)
			TGCACGCACATG	2.6	Intron	(96.1)
			GCGCCCCGCGCG	2.3	Intron	(33.7)
			GCACACGCACAT	2.0	Intron	(96.2)
			CGCGAGCACGCG	1.7	Intron	(30.0)
			CGCGGGCGCTGC	1.6	Intron	(30.4)
			AGGCACGCGCCT	1.6	Intron	(36.2)
			TGCACACACACG	1.5	Intron	(94.5)
Factor IX	38,059	HSFIXG	CATGCATGTGCACACA	2.4	Intron	(69.7)
			GCACGTGTGTGGGC	1.8	Intron	(15.3)
			CACACACGCATACACA	1.7	Intron	(92.7)
			CACATGCACACG	1.5	Promoter	(3.8)
			TGTGTGTATGCGTGTG	1.5	Intron	(22.5)
			GTGCACGTGCTT	1.5	Intron	(79.9)
Fc Receptor, Gamma Subunit	5,131	HSFCREB	CACACACACACACACA	2.4	Intron	(31.1)
			CACACACACACACACA	2.4	Intron	(31.2)
Gastrin	7,793	HSGASTA	GCGTGGGGGCGTGCGC	5.9	5' Flank	(28.7)
			CACGTGTGCGCT	1.7	Intron	(48.3)
			GCGTGTGCACAG	1.7	5' Flank	(3.2)
α-Globin Gene Cluster (Contains two genes)	12,847	HUMHBA4	GTCGGCGCGCACGC	14.9	Exon	(52.6)
			GTCGGCGCGCACGC	14.9	Exon	(82.2)
			TGCACGCGCACA	6.4	Exon	(55.1)
			TGCACGCGCACA	6.4	Exon	(84.8)
			CGTCCGGGTGCGCGCA	3.4	Promoter	(48.0)
			CGTCCGGGTGCGCGCA	3.4	Promoter	(77.6)
			CACACACACACACACA	2.4	5' Flank	(36.0)
			CGGGCGTGCCCCCGCG	1.6	Promoter	(51.4)
			CGGGCGTGCCCCCGCG	1.6	Promoter	(81.2)
			CAGGCGTGCGCC	1.5	5' Flank	(43.5)
			CGCACGTGGACG	1.3	Exon	(54.7)
			CGCACGTGGACG	1.3	Exon	(84.4)
			CCCGCGTGCACC	1.0	Promoter	(51.7)
			CCCGCGTGCACC	1.0	Promoter	(81.4)
β-Globin Gene Cluster (Region on Chromosome 11, containing 5 genes)	73,326	HUMHBB	TGTGTGCGCGCGTGTG	71.2	Intron	(48.5)
			GCACACACACACAC	3.0	3' Flank	(77.8)
			CACACACACACACACA	2.4	5' Flank	(42.9)
			TGTGTGTGTGTGTGTG	2.4	Intron	(48.5)
			TGTGTGTGTGTGTGTG	2.4	Intron	(55.2)
			GTGTGTGTGTGTGTGT	2.4	5' Flank	(81.1)
			CACACACATGTGTGCA	2.4	Flanking	(65.6)
			CATACGTGTGCACATG	2.0	Flanking	(35.9)
			ACACATGCACACGTATGT	2.0	Flanking	(99.2)
			TGTGTGTGTGCG	1.5	Flanking	(80.5)
			CACACATGCATGTG	1.3	5' Flank	(72.8)
			GTGTGTGTGTGTGT	1.3	Flanking	(79.8)
Theta-1-Globin	1,020	HUMGLTH1	GAGCGCGCGCGG	9.9	5' UTR	(13.0)
			CACCTGCACGCGTGCC	3.6	Exon	(61.1)
			CGCGCAGGCGCA	3.0	5' UTR	(22.8)
			GCGCTGGTGCGCGC	1.6	Exon	(29.9)
Gluco-cerebrosidase	7,604	HUMGCB1	ATAGGCGTGCGC	1.5	Intron	(30.1)
Glutathione Peroxidase	1,733	HUMGSHPXG	CGCTCGCGCGCA	3.5	5' UTR	(15.5)
			GTGCGGGGGCGCAC	1.6	Exon	(63.5)
Haptoglobin	11,551	HUMHPARS1	TGCATGTGTGTGTGTG	2.4	Intron	(21.1)
			CATGCATGTGTGTGTG	2.4	Intron	(21.4)
			ACAGGCGTGCGC	1.9	Intron	(14.4)
HSP* 90	8,210	HSHSP90B	GTGTGTACGCCCGC	2.2	Promoter	(2.7)
Histone H4	1,098	HSHIS4	TGGCGCGCGTCC	2.3	5' Flank	(10.1)
			GTGTACGCGCTC	1.7	Exon	(79.6)
Housekeeping Protein P3	4,379	HUMP3A	CCGCGCCCGCGC	4.3	Promoter	(19.1)
			GCGCGGGGGCGGACGC	1.5	Promoter	(7.7)
HMG-14*	8,882	HSHMG14A	GCCCGCACACGC	2.8	Intron	(12.2)
			TGTGTGTGTGTGTGTG	2.4	Intron	(32.5)
			TGTGTGTGTGTGTGTG	2.4	Intron	(54.4)
			GCGCCTGCGCAC	2.3	Promoter	(3.7)
			GCACGCGCCTTC	1.4	5' UTR	(11.3)
			GCACGCGCCTTC	1.4	5' UTR	(11.6)
			ACGTGCGTCCGC	1.1	Promoter	(6.4)
HMG-17*	7,195	HUMHMG17G	CTGCGCGCGCCT	7.0	Intron	(41.9)
			CTGCGCGCGCTT	6.3	Intron	(44.6)
			AAGCCCGCGCGC	3.3	5' UTR	(33.9)
			CCCGCGCACGAG	2.3	Intron	(41.3)
			TGCCCGCGGGCG	1.8	Intron	(37.5)
			TCGACGCGCGCC	1.5	Promoter	(30.2)
HPRT: Hypoxanthine Phosphoribosyltransferase	56,736	HSHPRTB	GCGTAGGCGCGCGGGC	6.4	Promoter	(2.6)
			GCCCGCGCGCCGGC	2.9	Promoter	(2.9)
			CACACACACACACACA	2.4	Intron	(39.1)
			TGTGTGTGTGTGTGTG	2.4	Intron	(39.2)
			TGTGTGTGTGTGTGTG	2.4	Intron	(58.7)
			CAGGCGCGTGCC	1.9	Intron	(23.2)
			CAGGCGCACGCC	1.9	Intron	(52.4)
			CGCGCGCCGGTA	1.6	Promoter	(2.3)
			CACACACACACACA	1.3	Intron	(59.0)
			TGTGTACACACACACA	1.3	Intron	(35.5)
			ACACACACGCATAC	1.0	Promoter	(1.1)
hst Protein	6,616	HSHST	GCGCGCACGCGGACAC	39.5	Exon	(43.3)
			GCGCGCACTGCT	1.4	5' UTR	(34.8)
Growth Hormone	2,657	HUMGHN	CGCCCGCGTGCAGG	2.8	Promoter	(6.9)
int-1 Oncogene (c-*myc*)	4,522	HUMINT1G	GCACGCACACGCGCGT	132	Exon	(75.6)
			CGCACGCGCACGCA	103	Exon	(73.9)
			GTGCGCACGTGC	16.8	Exon	(66.9)
			GCGGGCGCGCGT	12.4	Promoter	(0.3)
			TGCGTGCCCACGCACCTG	5.6	Intron	(45.5)
			CGTGCGCGAGTGCA	2.2	Exon	(31.9)
int-2 Oncogene	11,608	HUMINT2	GCGCGCGCCCGC	37.9	Intron	(20.8)
			CACGCGTGCCCGCGCC	18.1	5' Flank	(1.5)
			CGTGTGTGTGTGCGTG	13.4	3' Flank	(87.6)
			AAGAGCGCGCGC	6.1	Intron	(11.8)
			GTGGGCGCACGC	4.4	Intron	(73.8)
			GTGTGTGTGTGTGTGT	2.4	3' Flank	(88.9)
			CGCGGGGCGGGCGC	1.1	5' UTR	(7.9)
jun-B Oncogene	2,136	HSJUNCAA	CGCGCGCCTGGG	1.6	Exon	(60.8)
Kallikrein	6,139	HSKAL2	TGTGTGTGCATGTG	1.3	Intron	(39.3)
Keratin 18 (K18)	6,520	HUMKER18	TGTATGTGTGTGTGCA	1.0	Promoter	(24.4)
L-*Myc* Oncogene	7,011	HUMMYC3L	GCGCGCGCGCGG	129	5' UTR	(4.2)
			GCGCGCATGTGCGTGT	24.9	5' UTR	(4.9)
			CCCGGGCGCGCG	4.6	Promoter	(2.7)
			GCGCGCCCCCGCCC	2.1	Intron	(7.2)
			GGGCACGGGCGCGGGT	2.0	Intron	(6.4)
			GGGTCCGCGCGC	1.6	Intron	(8.5)

Table 4, Cont.

Name	Length	Locus	Sequence	Z-Score	Location	(%)
LYL-1 Protein	4,569	HUMLYL1B	CCGCCCGTGCGC	1.3	5' UTR	(13.6)
			TGACCGCACGCG	1.2	3' UTR	(86.8)
Meullerian Inhibiting Substance	3,100	HUMMIS	GTGCGCGCAGGCAC	6.5	Intron	(60.9)
			GCGCGCCTGCTCGCGC	3.2	Exon	(76.3)
			GGCCGTGCGCGC	2.7	Exon	(81.6)
			GCGCGCCCACCC	1.6	Exon	(87.5)
			CGCGCACGGCCC	1.3	3' Flank	(96.9)
MYCL2 Oncogene	3,854	HSMYCL2A	TGTGCGCGTGTGTG	12.5	5' UTR	(24.7)
			GCCCACGCGCCCCTGT	1.4	Promoter	(9.4)
β-Myosin Heavy Chain	28,438	HUMBMYH7	CGCATGCGCGGA	5.2	Promoter	(8.5)
			GTGTGTGTGTGTGC	3.0	Intron	(79.3)
			TGTGTGTGTGTGTGTG	2.4	Intron	(14.2)
			TGTGTGTGTGTGTGTG	2.4	Intron	(56.1)
			GTGTGTGTGTGTGTGT	2.4	Intron	(56.2)
			GCGTGCGCCTGT	1.9	Intron	(58.7)
			ACACGCACACACAGAC	1.8	Intron	(70.4)
			TGTGCATGTGTGCA	1.3	Promtoter	(4.6)
			GCGCGGGTGCGGGAGC	1.2	Exon	(85.5)
Opsin	6,953	HUMOPS	CACACACACACACACA	2.4	Intron	(29.7)
			CACACACACACACACA	2.4	Intron	(30.4)
			CTGCGCACGCCT	1.6	Exon	(7.1)
Ornithine Decarboxylase	8,841	HSSODB	CGGGCACGTGTGCG	5.5	Intron	(7.6)
			CACGTCCGCGCGGG	2.8	Promoter	(2.7)
			GCGGCCGCGCGC	2.0	Intron	(10.2)
Pancreatic Peptide	2,775	HUMPPPA	GCACGTGTGTGTGCAC	13.4	Intron	(42.1)
Perforin	6,218	HUMPRF1A	TGCGCACCTGCG	1.0	Exon	(83.1)
pim-1 Oncogene	6,113	HSPIM1A	GCGCACGGGCGTGC	13.8	Intron	(24.8)
			GCCCACCTGCGCGCCGCG	1.1	Exon	(19.6)
			CCCCGCGCGCCC	1.0	Promoter	(11.6)
PP14: Placental Protein 14	8,076	HSPP14B	TGTGTGTGTGTGTGTG	2.4	Promoter	(21.5)
			CACACACACACACACA	2.4	Promoter	(21.7)
			ACAGGCGCACGC	1.9	Intron	(72.8)
Prepro-Oxytocin-Neurophysin I	1,338	HSOTNPI	GCGCACCCGCAC	3.0	5' UTR	(30.2)
			GCCCGTACGCAC	1.0	Promoter	(19.5)
PCNA*	6,340	HSPCNA	GCGCGCGCTTGC	6.6	Promoter	(19.6)
Prostate Specific Antigen	7,130	HUMPSAA	CGTGTGTGCGCA	4.6	Exon	(59.4)
			AACACACGCACG	1.6	Intron	(35.7)
Protein C	11,725	HSPRCA	GCGCCCGCGCCC	4.3	Intron	(43.9)
			GCACCCGTGCGC	2.4	Exon	(44.7)
			GCGCGGGCGGGT	1.0	Intron	(41.7)
			ACCCGGGCGCGC	1.0	Intron	(42.9)
Prothrombin	20,801	HSTHB	CACACGCACACACA	3.8	Intron	(31.9)
			GTGCATGCACGC	2.6	Intron	(36.4)
			GCGTGTGCACAT	2.1	Intron	(67.9)
Pulmonary Surfactant Protein B (SP-B)	10,476	HUMSPBAA	CACACACACACACACA	2.4	Intron	(25.9)
			ACACACACACACACAC	2.4	Intron	(27.3)
			CACACACACACACACA	2.4	Intron	(28.5)
			CACACACACACACACA	2.4	Intron	(29.4)
			GCACACGCATAC	1.1	Intron	(38.0)
Pulmonary Surfactant Protein C (SP-C)	3,409	HSPLPSPC	TGTGCGCGCGCACATG	71.2	Intron	(20.5)
Pyruvate Dehydrogenase β-Subunit	8,872	HUMPDHBET	GCGGGCGTGCTCACGC	2.0	5' UTR	(26.8)
Ribosomal Protein S14	5,985	HSRPS14	CACACGTGGGTGTGCA	1.7	3' Flank	(99.1)
Ribosomal Protein S17	4,029	HSRPS17A	GCGCGTGTGCGA	5.1	Exon	(14.5)
Sex Hormone-Binding Globulin	6,087	HUMSHBGA	GTGCGTGCACCTGT	2.0	5' Flank	(30.2)
			CGCGCACCGGCACGCC	1.3	5' Flank	(16.5)
			CGCCCACACGCA	1.0	5' Flank	(42.2)
Somato-mammotropin	2,301	HSCS1	CGCGCGCACCAG	2.3	Exon	(41.4)
Thrombomodulin	4,593	HSTM	GCGAGCACGCGTGC	6.3	Exon	(32.1)
			CGTGCATGTGCG	3.7	Exon	(35.6)
			CGCCTGCACGCG	2.8	5' UTR	(15.1)
			TGCACGCGTGGG	2.7	Promoter	(5.6)
			GCACGTGCGGAC	1.2	Exon	(52.5)
			GCGCGAGGGCGC	1.1	Promoter	(8.8)
Thymidine Kinase	13,500	HSTKRA	CGCATGGGCGTGCG	4.6	Promoter	(1.2)
			GGGCGTGCGGGC	1.1	Intron	(86.9)
			GGGCGCACGTCC	1.0	Promoter	(2.3)
Thymidylate Synthase	18,596	HUMTS1	GCATGCAGGCGCGC	2.6	Exon	(6.3)
			CACAGGCGCGCG	2.6	Intron	(69.6)
			CATACGTGTGCATGTG	2.0	Intron	(33.4)
			GCGCACGCTCTC	1.4	Promoter	(4.3)
Tissue Factor	13,865	HSTFPB	CGTGCATGCATGCATG	5.2	Intron	(23.7)
			TGCATGTGCACACACA	2.4	Intron	(56.1)
			CACATGCATGCATGTG	2.4	Intron	(77.4)
			CACATGCATGCATGTG	2.4	Intron	(82.7)
			CACGTGTGTGTGTACA	2.1	Intron	(77.6)
			GGAGGCGCGCAC	1.8	Intron	(69.2)
			GCGCGCGGGGCACC	1.0	Promoter	(5.6)
			GCGGGGGCGGGCGC	1.0	Promoter	(5.2)
T-PA: Tissue Plasminogen Activator	36,594	HSTPA	GTGTGTATGCGTGTGC	3.3	Intron	(12.8)
			TGTGTGTGTGTGTGTG	2.4	Intron	(19.6)
			CACACACACACACACA	2.4	Intron	(46.2)
			GTGGGCACGTGC	1.2	Intron	(56.4)
Tryptase-I	2,609	HSTRYP1B	CGCGCCTACGCG	1.4	Exon	(19.2)
α-Tubulin	4,087	HUMTUBAG	CGTGTCTGTGCGCACG	2.9	Promoter	(7.5)
			GTGTGTGTGTGTGTGT	2.4	Intron	(20.5)
			TGTGTGTGTGTGTGTG	2.4	Intron	(20.9)
			GTGTGTGTGTGTGTGT	2.4	Intron	(22.0)
			GTGCACGCGTCT	2.0	Exon	(70.9)
			TGTGTGTGTGTGTG	1.3	Intron	(30.9)
β-Tubulin	3,284	HUMTUBBM	GCGCGCCCGCTC	3.5	5' UTR	(20.7)
			CCGTCCGCGCGC	1.6	Promoter	(9.5)
			CGGGCCCGCCCGCG	1.0	Intron	(32.9)
Ubiquitin-like Protein	3,583	HSUBILP	CGCGCGCGCCCG	37.9	Promoter	(14.1)
			ATGCGCGCGTGC	27.1	Promoter	(9.7)
			GCGGGCGCGCCC	4.9	Promoter	(12.8)
Urokinase-Plasminogen Activator	7,258	HUMUPAX	CACACACACGCTCACG	2.9	Intron	(29.1)

*Abbreviations used: ALP: Apolipoprotein; HSP: Heat Shock Protein; HMG: High mobility Group non-histone chromosomal protein; PCNA: Proliferating Cell Nuclear Antigen.