The EMBO Journal vol.5 no.10 pp.2737–2744, 1986

A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences

Pui S.Ho, Michael J.Ellison, Gary J.Quigley and Alexander Rich

Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Communicated by A.Rich

The ease with which a particular DNA segment adopts the left-handed Z-conformation depends largely on the sequence and on the degree of negative supercoiling to which it is subjected. We describe a computer program (Z-hunt) that is designed to search long sequences of naturally occurring DNA and retrieve those nucleotide combinations of up to 24 bp in length which show a strong propensity for Z-DNA formation. Incorporated into Z-hunt is a statistical mechanical model based on empirically determined energetic parameters for the B to Z transition accumulated to date. The Z-forming potential of a sequence is assessed by ranking its behavior as a function of negative superhelicity relative to the behavior of similar sized randomly generated nucleotide sequences assembled from over 80 000 combinations. The program makes it possible to compare directly the Z-forming potential of sequences with different base compositions and different sequence lengths. Using Z-hunt, we have analyzed the DNA sequences of the bacteriophage ϕX174, plasmid pBR322, the animal virus SV40 and the replicative form of the eukaryotic adenovirus-2. The results are compared with those previously obtained by others from experiments designed to locate Z-DNA forming regions in these sequences using probes which show specificity for the left-handed DNA conformation.

Key words: search strategies/sequence search/DNA conformation/DNA supercoiling

Introduction

It is now generally appreciated that local perturbations in DNA conformation arising as a consequence of negative supercoiling may act as signals to regulate a variety of genetic processes. The best characterized example of this class of structural alteration is the conversion of right-handed B-DNA to the left-handed Z-form (for review, see Rich *et al.*, 1984). From the crystal structures of oligonucleotides (Wang *et al.*, 1979), Z-DNA has been shown to be characterized by a regular alternation of *syn* and *anti* base conformations along each strand of the helix. Although the functional significance of Z-DNA *in vivo* is not entirely understood, there is now strong evidence for its involvement in genetic recombination (Kmiec *et al.*, 1985; Kmiec and Holloman, 1986). The central questions addressed in the present work concern which DNA sequences are most adept at becoming left-handed and how these sequences relate to the areas of major genetic activity in various genomes. Here we consider the DNA sequence alone, even though we are aware that the equilibrium between B- and Z-DNA may be strongly influenced by Z-DNA specific binding proteins (Lafer *et al.*, 1985).

For some time now, the operational paradigms directing Z-DNA formation have been: (i) that Z-DNA is formed most readily in alternating purine–pyrimidine sequences (Pohl and Jovin, 1972; Wang *et al.*, 1981; Drew and Dickerson, 1981; Jovin *et al.*, 1983); (ii) that the hierarchy for base pairs that can form Z-DNA is $d(GC)_n > d(CA)_n > d(AT)_n$ (Jovin *et al.*, 1983) and (iii) that longer stretches of these sequences form Z-DNA more easily than shorter stretches in negatively supercoiled DNA (Peck and Wang, 1983). These three rules have emerged from numerous biophysical studies on synthetic DNA polymers (Jovin *et al.*, 1983; Peck and Wang, 1983) and from the behavior of alternating purine–pyrimidine sequences cloned into closed circular plasmids (Singleton *et al.*, 1982; Peck and Wang, 1983; Haniford and Pulleyblank, 1983; Nordheim *et al.*, 1982). An algorithm based strictly on this set of rules had been used previously to predict the presence of Z-DNA in naturally occurring DNA sequences and these predictions have shown some correlation with published experimental results for these sequences (Konopka *et al.*, 1985). Currently, however, it is recognized that these simple rules do not entirely meet the requirements of the general situation. For example, while the alternating sequences $d(CG)_n$ and $d(CA)_n$ adopt the Z-conformation readily, the alternating sequence $d(TA)_n$ apparently does not (Jovin *et al.*, 1983; Quadrifoglio *et al.*, 1983; Haniford and Pulleyblank, 1985; Panyutin *et al.*, 1985; Greaves *et al.*, 1985; Ellison *et al.*, 1986a). The Z-forming potential of other, more complex combinations of alternating purine–pyrimidine sequences have not as yet been documented. Matters are further complicated by a variety of studies which have demonstrated either directly or indirectly that base pairs which disrupt the strict alternating character of a sequence can nonetheless be accommodated within the Z-DNA structural framework (Nordheim *et al.*, 1982; Wang *et al.*, 1985; Feigon *et al.*, 1985; Ellison *et al.*, 1985). From these observations, it is clear that, beyond a few limited cases, the Z-forming potentials of most types of DNA sequences are not immediately predictable.

The question of where Z-DNA is likely to form in naturally occurring DNA sequences has been addressed experimentally by several groups (DiCapua *et al.*, 1983; Nordheim and Rich, 1983; Stockton *et al.*, 1983; Revet *et al.*, 1984; Barton and Raphael, 1985; Hagen *et al.*, 1985; Miller *et al.*, 1983). The approaches used in these studies are conceptually similar; negatively supercoiled closed circular DNA is exposed to antibodies or chemical reagents that show specificity for Z-DNA, and the subsequent binding sites are qualitatively correlated with nearby stretches of purine–pyrimidine alternation. While these studies are useful, to date the resolution of these methods remains too low to establish unequivocably the correspondence between the probe binding site and its potential to form Z-DNA. Moreover, it has recently been suggested that polyclonal preparations of anti-Z-DNA antibodies can react with a non Z-DNA structure under certain conditions (Pulleyblank *et al.*, 1985). This emphasizes the point that many Z-DNA specific probes may not be entirely reliable. Re-

cently, chemical probes have been used that can define segments of Z-DNA formation at nucleotide resolution (Johnston and Rich, 1985; Herr, 1985).

In the present work, a thermodynamic approach is used in designing a computer program (Z-hunt) to search genomic sequences for regions most likely to adopt the Z-conformation. The closed circular genomes of the *Escherichia coli* virus ϕX174, the plasmid pBR322 and the mammalian tumor virus SV40 were selected for analysis by Z-hunt because previous empirical methods have been employed to locate their potential Z-DNA forming sequences (Revet *et al.*, 1984; DiCapua *et al.*, 1983; Barton and Raphael, 1985; Nordheim and Rich, 1983; Hagen *et al.*, 1985). The program treats DNA sequences as a collection of independent nearest neighbor interactions, each associated with its own free energy for the transition from B- to Z-DNA. Thus, it is possible to model the transition as a function of negative superhelicity by statistical mechanics. A number of the necessary energetic parameters for the 20 nearest neighbor interactions possible for Z-DNA have been, or are currently being, determined experimentally, giving this approach considerable predictive appeal (Peck and Wang, 1983; Ellison *et al.*, 1985, 1986a,b). The advantage of this particular search strategy is that the quantitative assessment of Z-forming potential is based not only on sequence composition but also on sequence length.

Results and Discussion

General features of Z-hunt

The ability of any given stretch of DNA to adopt the Z-form is directly related to the free energy required to bring about this conversion. Differences in the Z-forming potential of various DNA sequences can be thought of as arising from differences in the energies required to stabilize the various nearest neighbor interactions within each sequence in the Z-conformation, if it is assumed that longer range interactions are negligible. In B-DNA, there are 10 possible unique duplex nearest-neighbour interactions. In Z-DNA, however, there is potential for 20 such interactions since each base of the sequence can assume either the *syn* or the *anti* conformation. If the energies of all these interactions were known, a program can be envisaged which assigns an energetic value to each nucleotide in a sequence using a simple algorithm that is dependent only on the bases neighboring the nucleotides and whether they are *syn* or *anti*. Such a program could then locate stretches of nucleotides with high propensities to form Z-DNA by searching for local energy minima as it reads along the DNA sequence. However, there are two drawbacks to this approach. First, it would be particularly difficult to define the boundaries of many Z-forming stretches, since it is possible to propagate the left-handed conformation through minor energy maxima if this extension results in the bridging between good Z-forming regions (Ellison *et al.*, 1986a). Second, once the length of a Z-forming stretch is defined, sequences which differ in length cannot be directly compared because these energetic values do not provide information on the cooperative nature of the B to Z transition. We have thus opted for a search strategy which is not dependent on specifically defining the length of a Z-DNA forming stretch, but which still allows for comparisons between sequences having different lengths.

In closed circular DNA, the energy which is necessary to stabilize each interaction can be provided by the free energy associated with negative supercoiling. The degree of negative supercoiling that is sufficient to initiate the B to Z transition in a particular DNA segment may be used as an appropriate means by which the Z-forming potential of different types of DNA sequences can be compared. It has been demonstrated previously that the extent of Z-DNA formation as a function of negative superhelicity for the sequences $d(CG)_n$ and $d(CA)_n$ conforms nicely to a simple two-state statistical mechanical formulation of the zipper model (Peck and Wang, 1983; Vologodskii and Frank-Kamenetskii, 1984). Using this model, it has been possible to evaluate the free energies associated with the stabilization of several types of nearest neighbor interactions in the Z-form relative to the B-form of DNA (Table I). The computer program Z-hunt uses an expanded version of this two-state model which takes into consideration changes in sequence composition. By formulating a statistical mechanical treatment of the B to Z transition, all possible combinations of energy states for a given nucleotide sequence

Table I. Experimentally determined energetic parameters for dinucleotides associated with the B to Z transition

Nearest neighbor interactions		ΔG (kcal/mol) per dinucleotide	Reference
5′ *anti*−*syn* 3′	5′ *syn*−*anti* 3′		
5′-C G-3′ : : 3′-G C-5′	5′-G C-3′ : : 3′-C G-5′	0.66	Peck and Wang, 1983
5′-C A-3′ : : 3′-G T-5′	5′-A C-3′ : : 3′-T G-5′	1.34	Vologodski and Frank-Kamenetskii, 1984
5′-T A-3′ : : 3′-A T-5′	5′-A T-3′ : : 3′-T A-5′	2.4	Ellison *et al.*, 1986a
5′-C C-3′ : : 3′-G G-5′	5′-C C-3′ : : 3′-G G-5′	2.4	Ellison *et al.*, 1985
5′-C T-3′ : : 3′-G A-5′	5′-T C-3′ : : 3′-A G-5′	2.5	Ellison *et al.*, 1985
B−Z		5.0/junction	Peck and Wang, 1983
Z−Z junctions		4.0/junction	Ellison *et al.*, 1986b

Table II. Energy assignments for stabilizing dinucleotides in Z-DNA versus the B-DNA configuration[a]. The energies per dinucleotide are in kcal/mol

Dinucleotide	Conformation			
	AS	SA	(AS)[b]	(SA)[b]
CG	0.7[c]	4.0	4.0	4.0
GC	4.0	0.7[c]	4.0	4.0
CA	1.3[d]	4.6	4.5	4.5
AC	4.6	1.3[d]	4.5	4.5
TG	1.3[d]	4.6	4.5	4.5
GT	4.6	1.3[d]	4.5	4.5
TA	2.5[e]	5.9	5.6	5.6
AT	5.9	2.5[e]	5.6	5.6
CC	2.4	2.4	4.0[g]	4.0[g]
GG	2.4	2.4	4.0[g]	4.0[g]
CT	3.4[f]	3.4	6.3	6.3
TC	3.4	3.4[f]	6.3	6.3
GA	3.4	3.4[f]	6.3	6.3
AG	3.4[f]	3.4	6.3	6.3
AA	3.9	3.9	7.4	7.4
TT	3.9	3.9	7.4	7.4

[a]AS and SA represent *anti−syn* and *syn−anti* conformational phases of the dinucleotides listed.
(AS)[b] and (SA)[b] represent the conformation of dinucleotides which establish a Z−Z junction relative to the dinucleotide which precedes them (see text and Appendix). The free energy for each B−Z junction is set at 5.0 kcal/mol [10.0 kcal/mol total per segment (Peck and Wang, 1983)] and is assumed to be independent of base compositions. All values have been rounded off to the nearest 0.1 kcal/mol.
Superscript letters are associated with values which have been estimated experimentally and indicate references for that particular work.
[c]Peck and Wang, 1983.
[d]Vologodskii and Frank-Kamenetskii (1984)
[e]Ellison *et al.*, 1986a.
[f]Ellison *et al.*, 1985.
[g]Ellison *et al.*, 1986b.
All other values were estimated as described in the text.

are considered. Thus, it is not necessary to predetermine the actual bases involved in the transition since those bases having a high thermodynamic probability for existing in the Z conformation are heavily weighted in the calculations. In the present variation of the model, statistical weights have been included that account for the possible differences in the nearest neighbor interactions resulting from different possible base compositions and arrangements. This model and its underlying assumptions have been described elsewhere (Peck and Wang, 1983; Ellison *et al.*, 1985, 1986a) and are elaborated upon in the Appendix.

The program begins by subdividing the sequence to be examined into overlapping segments with lengths that can be varied by the user from 16 to 24 base pairs. The overlapping nature of this subdivision strategy reduces the possibility of dissecting a potentially good Z-forming candidate among two adjacent segments. Following the subdivision of the sequence, a second routine in Z-hunt analyzes each DNA segment and chooses for it that arrangement of dinucleotide free energy values for the B to Z transition from Table II which maximizes the alternating *syn−anti* character of the segment. This pre-screening of segments is designed to eliminate the most improbable Z-DNA energy states for a given segment, thereby simplifying the statistical mechanical analysis. The assignment of free energies to each segment is complicated by several practical and theoretical considerations which are described in greater detail below.

To assess the Z-forming potential of each segment, they are individually placed within the theoretical context of a closed circular plasmid molecule arbitrarily chosen to be the size of the plasmid pBR322 (4263 bp). Within this context, changes in the helical twist are only permitted to occur within the segment under investigation. Z-hunt then determines the average extent of Z formation by calculating the analytical solution to the partition function for the transition as a function of the negative superhelicity of the plasmid. The superhelical density that initiates on average the formation of one base pair per segment of Z-DNA has been selected as a measure of the onset of the transition. As discussed in the Appendix, this value was selected rather than the midpoint of the transition because of the search strategy and the practical limitations associated with the computations.

The calculated superhelical density could be used directly to compare the Z-forming potential of different DNA sequences; however, the meaning of such comparisons is not immediately obvious. As a normalized measure of the Z-forming potential for the segments, we have defined a more tangible quantity, 'Z-score', as the number of random nucleotide bases which must be searched on average to find a sequence which is as good or better at forming Z-DNA than the segment in question. These definitions and calculations are discussed in greater detail in the Appendix.

Energetic parameters for Z-hunt

The energetic parameters required for a statistical mechanical description of the B to Z transition in the sequences $d(CG)_n$ and $d(CA)_n$ as a function of negative superhelicity have been determined previously (Peck and Wang, 1983; Vologodskii and Frank-Kamenetskii, 1984). These values include the free energy required to stabilize the junctions that occur between B-DNA and Z-DNA as well as the free energies associated with the stabilization of the nearest neighbor interactions in Z-DNA. The energetic parameters which have been determined to date are listed in Table I. It should be emphasized that these experimental methods were incapable of resolving how the free energy changes are partitioned between a particular nearest neighbor interaction and the immediately adjacent interaction. As a result, the values listed in Table I refer to average values for the two types of interactions associated with each base pair.

The energetic parameters actually used in Z-hunt (Table II) have been assembled on the basis of two considerations. First, the program reads a sequence as blocks of dinucleotides to decrease the number of iterative calculations which must be performed to a manageable range. The propagation of a Z-DNA stretch by one dinucleotide block necessarily increases the free energy change associated with the transition by the sum of the nearest neighbor interactions in the block. Second, the free energy assignments to dinucleotides for which no experimental information is currently available were estimated using a simple algorithm. Each base contained within a dinucleotide is assigned a free energy value depending on whether it has a G or C base or a A or T base. In addition, this energy assignment is dependent on whether the base conformation is *anti* or *syn*. Since it is sterically more favorable for purine bases to adopt the *syn* conformation than pyrimidines (Haschemeyer and Rich, 1967), an additional energetic penalty is invoked when a pyrimidine is the *syn* conformation. These assignments have been estimated from the experimental values listed in Table I. The sum of these assignments for all dinucleotides are summarized in Table II and represent the estimated free energy change associated with the conversion of each dinucleotide from the B form to the Z form of DNA. The program can be modified accordingly as these values become known.

The assignment of base conformations and energetic values to dinucleotides within a given segment of DNA is further compli-

Table III. Test sequences with their corresponding conformational assignments and Z-scores

Sequence/conformation assignments	Z-score
1. C G C G C G C G C G C G C G C G C G C G C G A S A S A S A S A S A S A S A S A S A S A S	2×10^{11}
2. C G C G C G C G C G A S A S A S A S A S	4×10^{7}
3. C A C A C A C A C A C A C A C A C A C A C A A S A S A S A S A S A S A S A S A S A S A S	2×10^{5}
4. C A C A C A C A C A A S A S A S A S A S	2×10^{4}
5. C G C G C G C G C G G C G C G C G C G C A S A S A S A S A S A S A S A S A S A S	2×10^{8}
6. C G C G C G G C G C G C G C G C G C G C A S A S A S A S A S A S A S A S A S A S	7×10^{4}
7. C C C G C C C G C C C G C C C G C C C G C C C G A S A S A S A S A S A S A S A S A S A S A S	8×10^{4}
8. C A G G C A G G C A G G C A G G C A G G C A G G A S A S A S A S A S A S A S A S A S A S A S	1×10^{3}
9. C C C C C C C C C C C C C C C C C C C C C C C C A S A S A S A S A S A S A S A S A S A S A S	52
10. A T A T A T A T A T A T A T A T A T A T A T A T S A S A S A S A S A S A S A S A S A S A S A	38
11. A A A A A A A A A A A A A A A A A A A A A A A A A S A S A S A S A S A S A S A S A S A S A S	3×10^{-7}

Various test sequences are shown with their corresponding Z-scores as assigned by Z-hunt. Z-scores are defined as the number of random base pairs that must be scanned, on average, to find a sequence with equal or better Z-forming capacity relative to the sequence in question. The conformation selected by Z-hunt for each nucleotide (A for *anti* and S for *syn*) are indicated below each sequence. Bases which deviate from perfect purine–pyrimidine alternation are designated by dots above that nucleotide. Discontinuities in the conformational phases produced by Z–Z junctions are represented by gaps separating the sequence.

cated by an additional factor. Recent experiments have demonstrated that contiguous stretches of Z-DNA can exist in which the alternating *anti–syn* character of one stretch is out of phase with respect to the other. The interface between these two stretches has been termed a 'Z–Z junction' and has a free energy of formation estimated to be approximately 4.0 kcal/mol (Table I) (Ellison *et al.*, 1986b). To account for the possibility that such a phase change can occur within segments of DNA, an additional set of free energies was defined for each dinucleotide in which the energy for forming a Z–Z junction is added (Table II, columns 3 and 4). This energetic penalty for Z-formation is added when the conformational phasing of a particular dinucleotide changes with respect to that of the previous dinucleotide block. A more comprehensive treatment of Z–Z junctions and their structural consequences is presented below.

It is clear from the values listed in Table II and from the discussion above that a given segment of DNA can be assigned numerous combinations of conformations and associated energy values. To narrow the range of possibilities that Z-hunt must consider, the program selects only that arrangement of conformations and energies which maximizes the alternating *anti–syn* character of the sequence. This is directly related to minimizing the total energy required to stabilize the sequence in the Z-form. This minimization procedure is performed by a subroutine in the program which utilizes the propagation energies listed in Table II

Table IV. Sequences with the highest Z-DNA forming potential in ϕX174, pBR322, SV40 and Adeno-2 (RFI) genomic sequences

Genome	Rank	Sequence position	Sequence	Z-score
ϕX174	1	2233	G C G T G T A C G C G C A G G A	4×10^{5}
	2	1425	A T T T A T G C G C G C T T C G	5×10^{3}
	3	2226	T T A C C T T G C G T G T A C G	3×10^{3}
	4	994	T A C A C G C A G G A C G C T T	2×10^{3}
	5	3609	G C A A G A A C G C G T A C T T	2×10^{3}
	6	2289	A A A A A T T A C G T G C G G A	1×10^{3}
	7	242	T T C G T A T G C A G G G C G T	1×10^{3}
	8	4785	A T G C T T G G G A G C G T G C	9×10^{2}
	9	914	G T C G G G T A C G C A A T C G	7×10^{2}
	10	4793	G A G C G T G C T G G T G C T G	7×10^{2}
pBR322	1	1449	C A C G G G T G C G C A T G A T	6×10^{4}
	2	1410	G C C G C A C G C G G C G C A T	2×10^{4}
	3	258	C T A T G C G C A C C C G T T C	7×10^{3}
	4	2785	A G G G G T T C C G C G C A C A	5×10^{3}
	5	4250	A G G G G T T C C G C G C A C A	4×10^{3}
	6	4258	C G C G C A C A T T T C C C C G	4×10^{3}
	7	697	G G T G G G C G C G G G G C A T	2×10^{3}
	8	858	T T G C A C G C C C T C G C T C	2×10^{3}
	9	2074	C T C G C G C G T T T C G G T G	2×10^{3}
	10	2066	G C A G C T G C C T C G C G C G	1×10^{3}
SV40	1	4209	T A T G C C T G T G T G G A G T	8×10^{2}
	2	1289	G G G C A C A C C T A T G A T A	5×10^{2}
	3	4850	C T T T G C A C A C T C A G G T	3×10^{2}
	4	1281	T A T C A T T T G G G C A C C A	3×10^{2}
	5	2178	T A C A C A G G T G G G G A A A	3×10^{2}
	6	1986	G A G C T G C A G G G T G T G T	3×10^{2}
	7	553	T T C A G G T C C A T G G G T G	2×10^{2}
	8	1569	G T G C A A G T G C C A A A G C	2×10^{2}
	9	121	A T T G A G A T G C A T G C T T	2×10^{2}
	10	193	A T T G A G A T G C A T G C T T	2×10^{2}
Adeno-1	1	11474	T C C C G C G C G C G C A C A C	1×10^{6}
	2	16433	G C G C G T G C C C G T G C G C	1×10^{6}
	3	15970	C A G T G C G C G T G C G C G G	1×10^{6}
	4	12202	T G G C A C C C G C G C G C G C	9×10^{5}
	5	23505	A T C A A C G C G C G C G C A G	5×10^{5}
	6	22698	T G C G C G C G C G A G T T G C	4×10^{5}
	7	4089	G G G T T T T G C G C G C G C G	4×10^{5}
	8	11569	A C G T G C G C A C G C T T G T	2×10^{5}
	9	8650	C G G C G C C G C G C G C G G G	1×10^{5}
	10	6649	G C A A G C G C G C G C T C G T	1×10^{5}

Sequences are ranked in descending order by Z-scores. Z-scores are defined in the text and in the legend to Table III. The sequence position refers to the first nucleotide of each sequence according to the convention used in the National Biomedical Research Foundation Nucleic Acids Data Base.

to determine all the possible arrangements of *anti–syn* or *syn–anti* conformations for each dinucleotide in the sequence. It then searches for the arrangement associated with the lowest total energy for forming Z-DNA in that sequence. The individual energies for the dinucleotides from this minimum energy pathway are those used in the statistical mechanical calculation.

An examination of specific test sequences

Several of the capabilities of Z-hunt are effectively illustrated by an analysis of DNA sequences with simple, repeating compositions. The results from the analysis of these sequence are summarized in Table III. Most of these sequences incorporate nearest neighbor interactions whose free energies for the B to Z transition have been determined or estimated from previous experimental studies (sequences 1–7, 9 and 10). It is apparent from these analyses that sequences containing $d(CG)_n$ and $d(CA)_n$

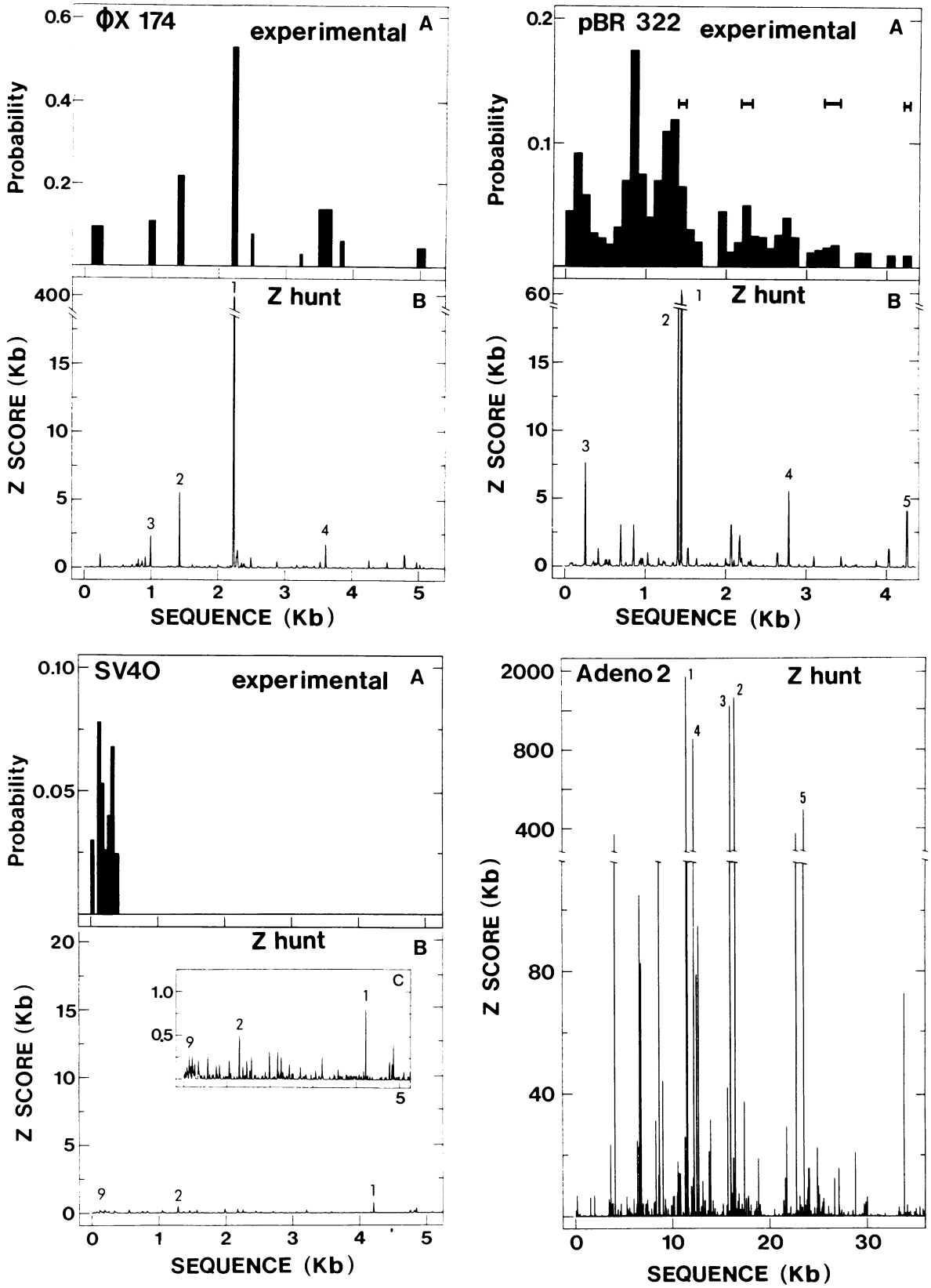

Fig. 1. Plots of potential Z-DNA sites versus sequence number in ϕX174, pBR322, SV40 and Adeno-2 (RFI). For the ϕX174 (Revet *et al.*, 1984), pBR322 (DiCapua *et al.*, 1983; Barton and Raphael, 1985), and SV40 (Hagen *et al.*, 1985), the top panel **(A)** represents the available experimental results from studies using Z-DNA specific probes. The results from anti-Z-DNA antibody studies (Revet *et al.*, 1984; DiCapua *et al.*, 1983; Hagen *et al.*, 1985) are plotted as the probability of observed binding events in electron micrographs versus sequence number. The cleavage sites in the plasmid pBR322 for the Z-DNA specific reagent CoDIP (see text) are indicated as bars, with the width of the bars representing the uncertainty associated with each site (Barton and Raphael, 1985). The results from the program Z-hunt for each genome are shown in the bottom panels. In all cases, the vertical axes are identical in scale. The insert **(c)** for SV40, however, is a 5-fold expansion along the *y*-axis.

show the greatest potential for forming Z-DNA. Stretches of $d(CG)_n$ are ranked notably higher than stretches of $d(CA)_n$ of the same length. This is entirely consistent with previous experimental findings (Peck and Wang, 1983; Haniford and Pulleyblank, 1983; Vologodskii and Frank-Kanenetskii, 1984).

The effect of sequence length on Z-DNA formation is evident by comparing $d(CG)_{12}$ with $d(CG)_6$ or $d(CA)_{12}$ with $d(CA)_6$. In both cases, the shorter sequences show a marked decrease in their Z-scores with respect to their longer counterparts.

A useful feature of Z-hunt is its ability to accommodate disruptions in the continuity of the *syn* and *anti* alternation that is characteristic of Z-DNA. The sequence $d(CG)_6(GC)_6$ (sequence 5), for example, contains a discontinuity (a Z−Z junction) which prevents the uniform propagation of the alternating *syn*−*anti* structure throughout the sequence. Similarly, $d(CG)_3(GC)_3(CG)_3$-$(GC)_3$ contains three such discontinuities. These disruptions in the alternating *syn*−*anti* character are accommodated in a given stretch of Z-DNA by changes from the *syn*−*anti* to an *anti*−*syn* phasing of the conformations, and vice versa. An energy minimized model of such a junction suggests that two adjacent guanine bases can assume the *syn* conformation and remain Watson−Crick paired with their complement cytidine bases in the *anti* conformation (Ellison *et al.*, 1986b). However, the validity of this model remains to be demonstrated. The effect of Z−Z junctions on the Z-forming abilities of DNA is apparent from the substantially reduced Z-scores of sequences 5 and 6 as compared with that for $d(CG)_{12}$. However, it is also clear that there is an energetic advantage afforded by splicing short, strong Z-stretches together into longer sequences via these junctions. The sequence $d(CG)_3(GC)_3(CG)_3(GC)_3$ with its three Z−Z junctions has a predicted Z-forming potential almost as great as $d(CA)_{12}$.

Recent studies have directly shown that base changes which disrupt the alternating purine−pyrimidine nature of a given sequence do not necessarily disrupt the alternation of *syn* and *anti* conformation inherent in the Z-DNA structure (Wang *et al.*, 1985; Feigon *et al.*, 1985; Ellison *et al.*, 1985). The energetics required to accommodate these changes suggest that certain types of non-alternating purine−pyrimidine arrangements of nucleotides may nonetheless have a propensity for Z-DNA formation. The ability of Z-hunt to search out such arrangements is illustrated by sequences 7−9 in Table III. Interestingly, while the sequence $d(CCCG)_6$ bears little resemblance to an alternating purine−pyrimidine sequence, its Z-score is only marginally lower than the Z-score for $d(CA)_{12}$. The Z-score for the poly(dC)poly(dG) sequence, however, is substantially worse. Interestingly, the sequence $d(TA)_{12}$, which is entirely alternating purine and pyrimidine, is predicted to be no better at forming Z-DNA than the poly(dC)poly(dG) sequence. This is a direct consequence of the unfavorable energy which is estimated for TA dinucleotides in the Z-conformation, as shown in Table II.

Potential Z-forming regions in genomic sequences

The detailed rules which comprise Z-hunt are broader than those which have been used in previous algorithms for predicting the potential occurrence of Z-DNA (Konopka *et al.*, 1985). As illustrated above, Z-hunt has the capacity to deal with conformational phase changes within sequences as a consequence of Z−Z junctions as well as the possibility that with purines and pyrimidines can adopt the *anti* and *syn* conformations, respectively. Furthermore, Z-hunt is capable of directly comparing the Z-forming potential of sequences with different lengths. Nevertheless, the program conforms generally to the previously held tenet that the best Z-forming sequences are likely to be long stretches of alternating purine−pyrimidine residues, coupled with a high percentage of GC base pairs (Konopka *et al.*, 1985). In the present work, Z-hunt has been used to search several genomic sequences for the presence of potentially strong Z-DNA forming regions. Since the presence of Z-DNA in the circular genomes φX174 (RFI) (Revet *et al.*, 1984), pBR322 (DiCapua *et al.*, 1983), and SV40 (Nordheim and Rich, 1983; Hagen *et al.*, 1985) has been previously detected with anti-Z-DNA antibodies and chemical probes (Barton and Raphael, 1985), the predictions of the program can be directly compared with accumulated empirical evidence for these three specific cases. Although the sequence of adenovirus-2 has not been analyzed experimentally, a search of its genome has also been included. The 10 sequences with the highest probability for Z-DNA formation according to Z-hunt are listed in Table IV. The general features of the analyses are best visualized by plotting the Z-score relative to the residue number for each segment of the sequence (Figure 1). To facilitate direct comparison of these results with the available experimental results, the frequency of antibody binding to various regions of each sequence are displayed on the same sequence scale and directly over the complementary Z-hunt results.

The predictions from Z-hunt for φX174 are in remarkably close agreement with the results obtained from antibody binding experiments (Revet *et al.*, 1984). There are strong correlations between the results from the program and the experimental analyses in both the positions and the relative probabilities for segments forming Z-DNA.

While there are clear correlations between the program and empirical analyses (DiCapua *et al.*, 1983; Barton and Raphael, 1985) for the plasmid pBR322, there are some notable discrepancies. The strongest segments of Z-DNA formation predicted by the program begin at positions 1410 and 1460 of the pBR322 map. While this region appears to be a major site for antibody reactivity, a stronger antibody binding site appears at position 800−900. Interestingly, Z-hunt ranked this particular region ninth. The Z-DNA specific chemical probe cobalt(III) tris(diphenylphenanthroline) (CoDIP) entirely failed to detect this site. Overall, there is only marginal correlation between the sites recognized by CoDIP with either the antibody binding studies or the Z-hunt analysis of the pBR322 plasmid. The only sites clearly in common with all three analytical methods lie between 1400 and 1500 and contain the sequences d(CACGGGTGCGCATG) and d(CGC-ACGCG). The lack of correspondence between the results from CoDIP and the results presented here and by others is puzzling. For example, CoDIP is reported to cleave the sequence d(GTA-TATATGAGTA) located at position 3265 on the pBR322 map. In contrast, it has been found that a $d(TA)_4$ segment flanked in phase by two strong Z-forming $d(CG)_6$ sequences fails to adopt the Z-conformation when contained in a negatively supercoiled plasmid, even though the $d(CG)_6$ segments have flipped to the left-handed form (Ellison *et al.*, 1986a). The apparent reluctance of $d(TA)_4$ to adopt the Z-conformation under conditions similar to those reported for the CoDIP studies raises questions concerning the relative specificity of CoDIP for Z-DNA over other possible altered DNA conformations.

The most striking contrast between the program predictions and the published experimental results arises from the analysis of the SV40 virus. When the Z-scores for this genome are plotted on the identical scale to those for φX174 and pBR322, the SV40 genome appears devoid of sequences with strong Z-forming potential. The sequence with the highest Z-score is at best two orders of magnitude worse than the best Z-DNA sequence from all the other genomes analyzed, and none of the five highest ranked SV40 sequences appear to act as Z-DNA antibody determinants.

The antibody, however, has been shown to map Z-DNA to a region between residues 1 and 400 (Nordheim and Rich, 1983; Hagen *et al.*, 1985). This region includes two identical eight base pairs alternating purine−pyrimidine stretches centered at the two *Sph*1 restriction sites of the viral enhancer. Both these sequences are ranked ninth by the program. The difficulty in this comparison lies in our inability to understand the behavior of poor Z-DNA forming sequences, particularly when these are subjected to very high levels of negative supercoiling, as used in at least one (Nordheim and Rich, 1983) set of antibody binding experiments.

The presence of sequences with strong Z-forming potential in Adeno-2 can probably be attributed to two factors. Adeno-2 is approximately seven times larger than other sequences examined in this study, thereby increasing the probability of detecting infrequently occurring sequences with strong propensity for the formation of Z-DNA. Furthermore, the four sequences in Adeno-2 with the highest Z-scores are all located in a 10 kb region where the percent of d(CpG) and d(GpC) dinucleotides is greater than 11%. These dinucleotides represent only 3% of the SV40 genome. Given the bias of the model for these particular dinucleotides, it is not unreasonable to attribute the dramatic difference observed between Adeno-2 and the SV40 genomes to this difference in their base compositions.

There are several straightforward explanations that can account for the discrepancies between the Z-hunt analyses and the experimental measurements. In all of the cited experiments involving the use of anti-Z-DNA antibodies, polyclonal preparations of the antibodies were raised against the polymer d(CG)·d(GC) modified to exist in the Z-form under physiological conditions. Various studies have shown that these preparations show selective reactivity towards $d(CG)_n$ over $d(CA)_n$ when both of these sequences are in the Z-conformation (Zarling *et al.*, 1984). Therefore, in certain instances, the degree of antibody binding to a given stretch of Z-DNA may reflect the affinity of the antibody for a given sequence rather than the ease with which that sequence adopts the Z-form. In addition, the affinity of the antibody will depend heavily on whether or not bivalent binding of the arms of the antibody is favored in one stretch of DNA as compared with another. Matters are further complicated by previous reports that certain preparations of anti-Z-DNA antibodies show a high enough affinity to some sequences to measurably shift the B-DNA to the Z-DNA equilibrium in favor of Z-formation (Lafer *et al.*, 1985). Thus sequences which exist in the Z-form relatively infrequently may still serve as strong antigenic determinants. It is also possible that polyclonal preparations of anti-Z-DNA antiodies contain minor species of antibody that recognize alternative non Z-DNA conformations. Pulleyblank *et al.* (1985) have observed that the negative supercoiling-induced structural transition in poly d(GA)·d(CT) reacts with polyclonal preparations of anti-Z-DNA antibodies at pH 6.0. It is therefore conceivable that certain minor sites of antibody reactivity may be due to some other type of structural transition in DNA.

In view of the arguments presented above, it is reasonable to believe that some of the differences between the Z-hunt predictions and the antibody results arise from uncertainties associated with the specificity of antibody binding. Practical and theoretical limitations inherent in the design of Z-hunt could also account for some of these disrepancies. While several of the energetic rules used in Z-hunt are based on experimentally derived values, many more of these parameters have only been estimated using the simple algorithm described above. In particular, we have assumed in Z-hunt that the free energies of formation of B−Z and Z−Z junctions are independent of sequence composition; this may not be entirely valid. Until more becomes known about the energetics for the formation of these junctions, the treatment of these parameters by Z-hunt remains uncertain. Furthermore, even those values which were determined experimentally may remain valid only within narrow ranges of pH, ionic strength and temperature. The effect of environmental conditions on these energy parameters is not well understood (Rich *et al.*, 1984). In addition to these practical limitations, it is important to emphasize that Z-hunt is based upon a simple two-state statistical mechanical model in which base pairs can only exist in either the B-form or the Z-form. It is currently unclear how the existence of other possible underwound forms of DNA will influence the energetics of the B to Z transition in certain types of DNA sequences.

In spite of the limitations to the program cited above, the ability of Z-hunt to assess and compare the potential for Z-formation in sequences of varying length, as well as its consideration of structural peculiarities such as changes in conformational phasing, establishes this approach as the most comprehensive search strategy to date. Furthermore, the program can be expanded or modified to include experimentally determined energetic parameters for the B to Z transition as well as other competing structural transitions as these values become available. The application of analytical approaches such as Z-hunt may ultimately prove useful for understanding the relationship between the structural versatility of supercoiled DNA and its role in biologically important processes.

Acknowledgements

We acknowledge the assistance of Raymond Kelleher, James K.Prater and L.Bridges for help in the formulation of the Z-hunt program. This research was supported by grants from the National Institutes of Health, the American Cancer Society, the National Aeronautics and Space Administration, the Office of Naval Research and the National Science Foundation. P.S.H. is a fellow of the American Cancer Society and M.J.E. was supported by a fellowship from the Medical Research Council of Canada.

References

Barton,J.K. and Raphael,A.L. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 6460−6464.

Depew,R.E. and Wang,J.C. (1975) *Proc. Natl. Acad. Sci. USA*, **72**, 4265−4279.

DiCapua,E., Stasiak,A., Koller,T., Brahms,S., Thomae,R. and Pohl,F.M. (1983) *EMBO J.*, **2**, 1531−1535.

Drew,H.R. and Dickerson,R.E. (1981) *J. Mol. Biol.*, **151**, 535−556.

Ellison,M.J., Kelleher,R.J. III, Wang,A.H.-J., Habener,J.F. and Rich,A. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 8320−8325.

Ellison,M.J., Feigon,J., Kelleher,R.J. III, Wang,A.H.-J., Habener,J.F. and Rich, A. (1986a) *Biochemistry*, **25**, 3648−3655.

Ellison,M.J., Quigley,G.J., Johnston,B., Kelleher,R.J. III, Wang,A.H.-J. and Rich,A. (1986b), in preparation.

Feigon,J., Wang,A.H.-J., van der Marel,G.A., van Boom,J.H. and Rich,A. (1985) *Science*, **230**, 82−84.

Greaves,D.R., Patient,R.K. and Lilley,D.M.J. (1985) *J. Mol. Biol.*, **185**, 461−478.

Hagen,F.K., Zarling,D.A. and Jovin,T.M. (1985) *EMBO J.*, **4**, 837−844.

Haniford,D.B. and Pulleyblank,D.E. (1983) *Nature*, **302**, 632−634.

Haniford,D.B. and Pulleyblank,D.E. (1985) *Nucleic Acids Res.*, **13**, 4343−4363.

Haschemeyer,A.E.V. and Rich,A. (1967) *J. Mol. Biol.*, **27**, 369−384.

Herr,W. (1985) *Proc. Natl. Acad. Sci. USA*, **85**, 8009−8013.

Johnston,B. and Rich,A. (1985) *Cell*, **42**, 713−724.

Jovin,T.M., McIntosh,L.P., Arndt-Jovin,D.J., Zarling,D.A., Robert-Nicoud,M., van de Sande,J.H., Jorgensen,K.F. and Eckstein,F. (1983) *J. Biomol. Struct. Dynam.*, **1**, 21−57.

Kmiec,E.B. and Holloman,W.K. (1986) *Cell*, **44**, 545−554.

Kmiec,E.B., Angelides,K.J. and Holloman,W.K. (1985) *Cell*, **40**, 139−145.

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

Lafer,E., Sousa,R. and Rich,A. (1985) *EMBO J.*, **4**, 3655−3660.

Miller,F.D., Jorgenson,K.F., Winkfein,R.J., van de Sande,J.H., Zarling,D.A., Stockton,J. and Rattner,J.B. (1983) *J. Biomol. Struct. Dynam.*, **1**, 611−620.

Nordheim,A. and Rich,A. (1983) *Nature*, **303**, 674–679.
Nordheim,A., Lafer,E.M., Peck,L.J., Wang,J.C., Stollar,B.D. and Rich,A. (1982) *Cell*, **31**, 309–318.
Panyutin,I., Lyamichev,V. and Mirkin,S. (1985) *J. Biomol. Struct. Dynam.*, **2**, 1221–1233.
Peck,L.J. and Wang,J.C. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 6206–6210.
Pohl,F.M. and Jovin,T.M. (1972) *J. Mol. Biol.*, **57**, 375–395.
Pulleyblank,D.E., Shure,M., Tang,D., Vinograd,J. and Vosberg,H.-P. (1975) *Proc. Natl. Acad. Sci. USA*, **72**, 4280–4284.
Pulleyblank,D.E., Haniford,D.B. and Morgan,A.R. (1985) *Cell*, **42**, 271–280.
Quadrifoglio,F., Mann,G., Yathindra,N. and Crea,A. (1983) In Pullman,B. and Jortner,J. (eds), *Nucleic Acids: The Vectors of Life*. D.Reidel, Dordrecht, Holland, pp. 61–74.
Revet,B., Zarling,D.A., Jovin,T.M. and Delain,E. (1984) *EMBO J.*, **3**, 3353–3358.
Rich,A., Nordheim,A. and Wang,A.H.-J. (1984) *Annu. Rev. Biochem.*, **53**, 791–846.
Singleton,C.K., Klysik,J., Stirdivant,S.M. and Wells,R.D. (1982) *Nature*, **299**, 312–316.
Stockton,J., Miller,F.D., Jorgenson,K.F., Zarling,D.A., Morgan,A.R., Rattner, J.B. and van de Sande,J.H. (1983) *EMBO J.*, **2**, 2123–2128.
Vologodskii,A.V. and Frank-Kamenetskii,M.D. (1984) *J. Biomol. Struct. Dynam.*, **1**, 1325–1333.
Wang,A.H.-J., Quigley,G.J., Kolpak,F.J., Crawford,J.L., van Boom,J.H., van der Marel,G. and Rich,A. (1979) *Nature*, **282**, 680–686.
Wang,A.H.-J., Quigley,G.J., Kolpak,F.J., van der Marel,G., van Boom,J.H. and Rich,A. (1981) *Science*, **211**, 171–176.
Wang,A.H.-J., Gessner,R.V., van der Marel,G.A., van Boom,J.H. and Rich,A. (1975) *Proc. Natl. Acad. Sci. USA*, **82**, 3611–3615.
Zarling,D.A., Arndt-Jovin,D.J., Robert-Nicoud,M., McIntosh,L.P., Thomae,R. and Jovin,T.M. (1984) *J. Mol. Biol.*, **176**, 369–415.

Received on 16 May 1986

Appendix

Assignment of energies for the B to Z transition in dinucleotides

Z-hunt is based upon a two-state statistical model that evaluates and compares the degree of Z-DNA formation as a function of negative superhelicity in different segments of DNA. In this model, the dinucleotides contained in a particular stretch of DNA are permitted to occupy either the right-handed B state or one of the two phases for the left-handed Z-state (*syn–anti* or *anti–syn*). Associated with each one of these left-handed states is a free energy difference for the conversion of a dinucleotide from B to Z-DNA (ΔG_{B-Z}). The value of (ΔG_{B-Z}) selected for dinucleotides in each of these two states depends upon the sequence of the dinucleotide and whether or not the particular dinucleotide maintains the *syn* and *anti* alternation of the dinucleotides to either side of it. Tables of (ΔG_{B-Z}) values for each of the 16 dinucleotides in each of the allowable conformational states are listed in Table II. Some of these values have been determined empirically (Table I) while others have been estimated using the algorithm described in the text (Table II).

The conformational fate of each dinucleotide within a given DNA stretch is determined by a subroutine of Z-hunt called Spawn. Spawn selects that arrangement of dinucleotide conformations from Table II that minimizes the total free energy for the B to Z transition. This subroutine serves to select the most energetically favored conformation of the sequence, thereby reducing the number of microstates that must be considered by the statistical mechanical analysis described below. In view of the high energetic penalty associated with dinucleotide conformations which form Z–Z junctions, Spawn has the effect of maximizing the alternating *syn* and *anti* character of a given sequence in the Z-form.

Statistical mechanical treatment of conformationally assigned DNA segments

Determining the extent of Z-DNA formation as a function of negative supercoiling involves placing individual segments of DNA, whose energetics for the transition were assigned by Spawn, within the context of a closed circular duplex DNA molecule. The extent of Z-DNA formation in each segment is monitored as the change in twist of the segment from the right-handed form to the left-handed form as a function of the linking diference using a statistical mechanical formulation of the zipper model. It has been demonstrated previously that a simple two-state derivation of the zipper model effectively describes the B to Z transition in $d(CG)_n$ and $d(CA)_n$ as functions of negative supercoiling (Peck and Wang, 1983; Vologodskii and Frank-Kamenetskii, 1984). We have extended this model to account for differences in sequence composition (Ellison *et al.*, 1986a). In this model, it is assumed that each microstate involves only one nucleation event per DNA segment and that the free energy of nucleation is independent of sequence composition. The configuration partition function for a segment of DNA composed of n non-identical dinucleotides is given by the series expansion:

$$Q = 1 + \sum_{i=1}^{n}\sum_{k=1}^{n}\sigma\left(\prod_{j=i}^{k} S_j\right)\exp\{(-K/RT)[\alpha-\alpha_0^0 - (\sum_{j=i}^{k} a_j) - 2b]^2\}$$

Here S_j is the equilibrium constant for the transition of the jth dinucleotide in the segment from the B-state to its previously assigned left-handed state. S_j takes the form $\exp(-\Delta G_{B-Z}/RT)$. The a_j term denotes the change in the twist associated with the conversion of a dinucleotide from the right-handed form to the left-handed form ($a_j = 0.18$ turns). The parameter b is the degree of unwinding occurring at the B–Z junction ($b = 0.4$ turns). σ is the nucleation parameter which includes the free energy change associated with the formation of two B–Z junctions and takes the form $\exp(-10$ kcal$/RT)$. The experimental determinations of a_j, b and σ have been reported previously (Peck and Wang, 1983). K is the proportionality constant for the free energy of supercoiling (Pulleyblank *et al.*, 1975; Depew and Wang, 1975) and is taken to be 1100 RT/N, where N is the size of the closed circular DNA in base pairs. In Z-hunt, N is fixed at the size of the pBR322 plasmid (4363 bp). The quantity $\alpha-\alpha_0^0$ represents the change in linkage of the closed circular DNA molecule from the theoretically relaxed configuration.

The average value of the change in twist $<\Delta Tw>$ (an experimentally measured quantity) as a function of the plasmid linking difference is expressed as a product over the probability sum (Ellison *et al.*, 1986a):

$$<\Delta Tw> = Q^{-1}\{\sum_{i=1}^{n}\sum_{k=i}^{n}[(\sum_{j=i}^{k} a_j)+b]\sigma(\prod_{j=i}^{k} s_j)\exp[(-K/RT)\times$$

$$(\alpha-\alpha_0^0-(\sum_{j=i}^{k} a_j)-2b)^2]\}$$

Using the above relationship, Z-hunt numerically evaluates $-<\Delta Tw>$ as a function of the linking difference, $\alpha-\alpha_0^0$ at which $-<\Delta Tw> = 1.0$ turn (an average of 1 bp of Z-DNA per segment).

The program could alternatively have defined the formation of Z-DNA in a segment as the $<\Delta Tw>$ at the midpoint of the transition. This would, however, overweight the contribution of poor Z-forming dinucleotides in many segments, thereby reducing the apparent potential of good Z-forming sequences in those segments. Furthermore, locating the midpoint of the transition would require extending the calculations to values of superhelical densities at which the transition asymtopes. This would significantly increase the analysis time for each segment. At $<\Delta Tw> = 1.0$ turn, the partition function is dominated by the energies of the dinucleotides with the strongest propensity to form Z-DNA in the segment. This allows the program to evaluate the best Z-forming sequences within each segment with a minimum of calculations. However, it should be noted that with this search strategy, sequences which begin to undergo the transition at low superhelical densities, but are not highly cooperative, will be ranked higher than those whose transition is more cooperative, but initiate at a higher superhelical density. These values form the basis of comparison for the Z-forming potential (which we term the Z-score) of different segments of DNA.

To calculate the Z-score for a particular segment, 80 000 randomly generated segments were analyzed for their Z-DNA forming potential using Z-hunt. The cumulative results approximate a normal distribution along with its associated mean and SD values for superhelical densities required to initiate formation of Z-DNA within an 'average' sequence. The Z-score for a given sequence is thus defined as the deviation, in standard units, of its calculated superhelical density from the mean of the normal distribution. The Z-score is converted to a numerical probability through analytical integration of the normal distribution up to the Z-score. By multiplying this probability by the size of the segment, we determine the number of random nucleotide bases which must be searched on average to find a sequence which is as good or better at forming Z-DNA than the segment in question.

The program Z-hunt

The program is written in VMS FORTRAN for a VAX model 11-780 computer. On a system equipped with 4 Mbytes memory, the total c.p.u. time required for a complete analysis of a 5000 base pair sequence is approximately 1.5 h. There are two formats currently available for the input of a sequence. The first is identical to the output format from the nucleic acids sequence analysis program RUNSEQ and the second is free format, in which an input file is headed by two lines of descriptive text followed by the sequence as a continuous character string with no more than 80 characters per line. One output file format of Z-hunt lists the 10 (or any number specified by the user) segments of the sequence with the highest propensity to form Z-DNA. In this file are included the phase assignments for the dinucleotides, the starting and ending sequence numbers of the segment, and the Z-score in kb for the segment. The second output file lists sequentially the Z-scores of all segments with their sequence numbers in a form which can be plotted (as in Figure 1). The program is available on request.