A consensus sequence for cleavage by vertebrate DNA topoisomerase II

Jeffrey R.Spitzner and Mark T.Muller*

The Ohio State University, Department of Molecular Genetics, Columbus, OH 43210, USA

Received December 30, 1987; Revised and Accepted April 29, 1988

ABSTRACT

Topoisomerase II, purified from chicken erythrocytes, was reacted with a large number of different DNA fragments and cleavages were catalogued in the presence and absence of drugs that stabilize the cleavage intermediate. Cleavages were sequenced to derive a consensus for topoisomerase II that predicts catalytic sites. The consensus is:

	-10		-8		-6		-4		-2	-1		1		3		5		7		
5'	A/G	N	T/C	N	N	C	N	N	G	T/C	^	N	G	G/T	T	N	T/C	N	T/C	3'

where N is any base and cleavage occurs at the indicated mark between -1 and + 1. The consensus accurately predicts topoisomerase II sites in vitro. This consensus is not closely related to the Drosophila consensus sequence, but the two enzymes show some similarities in site recognition. Topoisomerase II purified from human placenta cleaves DNA sites that are essentially identical to the chicken enzyme, suggesting that vertebrate type II enzymes share a common catalytic sequence. Both viral and tissue specific enhancers contain sites sharing strong homology to the consensus and endogenous topoisomerase II recognizes some of these sites in vivo.

INTRODUCTION

Eukaryotic type II topoisomerases have been purified from a number of sources (1-5, Muller et al., manuscript submitted) and the activities have been characterized in vitro (for reviews, 6,7). Catalytic sites of topoisomerase II (topo II^a) on DNA can be detected by uncoupling the breakage and rejoining activity using detergents to denature the enzyme, thereby yielding DNA molecules with double strand breaks which can be sequenced (8,9); we refer to this procedure as a topo II cleavage reaction. The efficiency of trapping the cleavage intermediate is significantly increased through the use of drugs that stabilize the half life of the covalent complex between topoisomerase and DNA (10,11), and these agents have been particularly useful for correlating in vitro sites with sites in chromatin that are mediated by endogenous topoisomerase (12-15). Several studies support the idea that the in vitro cleavages accurately reflect the physiological activity of topoisomerase II (12-14). For this reason, the acquisition of new information about this enzyme would be facilitated by the ability to predict sites which display the features of a topo II catalytic site.

In yeast, topoisomerase II is an essential gene product required for completion of mitosis (16-18). In accordance with this cellular role, topoisomerase II levels are under cell cycle regulation, and this observation has opened up new avenues for anti-tumor drug development (19,20). Topoisomerase II is a major component of the nuclear scaffold and matrix (21-23) and matrix associated regions of DNA contain topoisomerase II binding sites (24). The exact role of topoisomerase II in the nuclear matrix is unknown, but this enzyme may be important in regulating DNA topology of chromatin domains, or in propagating alternate DNA structures in chromatin, as suggested from the co-localization of topoisomerase II sites with a subset of DNase hypersensitive regions (13,15). The implication that topoisomerase II is involved in pivotal aspects of gene control is reasonable. While it is likely that topo II performs similar cellular functions in all eukaryotic cells, some provocative conclusions in a mammalian system were drawn by comparison of sequences to a consensus derived for an invertebrate (Drosophila melanogaster) topo II (24). In considering the significance of these findings, we felt that it would be valuable to determine a consensus recognition sequence applicable to vertebrate type II DNA topoisomerases.

RESULTS

A consensus for topoisomerase II cleavage of DNA

Chicken topoisomerase II cleavage reactions were performed on a wide range of DNA substrates to minimize bias introduced by differences in base composition. These variations would obviously not be represented in a single plasmid or gene, despite the fact that a large number of different cleavage sites could be conveniently compiled. DNA substrates were 5' end labeled fragments of DNAs from a variety of organisms and viruses (Table 1). Reactions were carried out in the presence and absence of the topoisomerase II inhibitor m-AMSA (11) and were analyzed on sequencing gels with chemical sequence ladders (25) to resolve cleavages to single residues. When possible, both top and bottom strands were analyzed (see Table 1). Relatively strong cleavages, selected as the most intense bands on the gel, were compiled with up to 50 bp flanking the cleavage site in both 5' and 3' directions; however, in some fragments this was not possible if a strong site was close to the end of the fragment (in this case "N"s were introduced; see statistical methods). Sequence data were centered with respect to the cleavage site and stored in a computer data base.

It is clear that although topoisomerase II is a homodimer, the recognition sites are not necessarily symmetric, but instead possess a polarity of (arbitrarily chosen) top strand and bottom strand. It was necessary to align the sites as "top strand" cleavages by reading 5' to 3' on the top strand or reading 5' to 3' on the bottom strand. This distinction must take into account the 5' extension of 4 bp (8,9, Muller et al., submitted), which changes the position of cleavage relative to nucleotide positions as assigned in the data base. Alignment was a computer assisted iterative process: Certain nucleotide preferences were noted for positions flanking the cleavage sites and the program determined which strand presented the best match to these preferences. That strand was then aligned as "top strand" and the process repeated. Finally a new modal (or consensus) cleavage pattern was defined by examining nucleotide percentages at each position with respect to the cleavage site and selecting all positions with single nucleotide percentages greater than 50% or with two nucleotides combined percentages greater than 70%. Sequences were then aligned with this pattern until a consensus was approximated which gave the highest overall percentage match to the 71 sequences in the data base (69.6% not counting matches to "N" bases) compared to a data base of random sequences generated by the computer. This degenerate consensus chicken topoisomerase II cleavage site (top strand) is shown in Table 2 below the nucleotide proportions at each position for the entire data base.

Nucleotide distribution is essentially random outside the consensus region

We combined all cleavages to calculate the redundancy (or deviation from randomness) of the composite data at each position relative to the cleavage site (46). This analysis revealed which residues are most likely to be important in defining a consensus. The percent redundancy at each position is shown graphically in Figure 1E. There was significant redundancy at the consensus positions compared to random base proportions (i.e. percent redundancy near 0) at the positions where an N is placed in the consensus and at positions outside of the consensus region. The exceptions are positions 10 and 11 (bases 3' of the cleavage site) in which the pyrimidines dominate sufficiently to show redundancy; however, inserting a pyrimidine at either position decreased the percentage match of the consensus to the real data base and increased the match to a data base of random sequences.

Figure 1A-D shows the distribution of selected species at each position in the data base. In each case, the proportions differ greatly from the mean (of the real data base) only within the consensus region; thus, sequences more than 8 bases downstream of the cleavage site have an essentially random distribution of bases. Collectively, the data indicate that topoisomerase II makes its base-specific interactions with DNA in the 18 base consensus region; however, this does not preclude protein DNA interactions in flanking regions. Indeed, DNase I footprinting experiments suggest that topoisomerase II protects a region wider than the 18 bp consensus (data not shown).

Consensus is the major determinant for DNA cleavage by topoisomerase II

We conducted a variety of analyses to investigate the possibility that factors other than a direct match to consensus play roles in determining topoisomerase II cleavages. The di- and trinucleotide frequencies at each position were calculated (data not shown) and were found only to reflect the frequencies suggested by the consensus; thus, GC and GT were the most common dinucleotides found at positions -2, -1, but other dinucleotides were also found. A search for palindromes was performed to investigate whether topoisomerase II recognizes specific symmetry between the two strands in individual cleavages. A computer algorithm was developed which compares the 5' to 3' top strand sequence with the 5' to 3' bottom strand

Table 1. DNA substrates used to derive consensus sequence

Sequence Number:	DNA Source:[a]	Restriction sites (size)[position]:[b]	Comments on the sequence (ref):
1.	kDNA	*StuI-HinfI (226 bp) [2515-2289]	Bent DNA fragment (see footnote c for source of sequence)
2.	"	*StuI-HinfI (549) [1-549]	footnote c
3.	"	*AvaII-SacII (258) [1763-2020]	"
4.	"	*HinfI-SacII (263) [2283-2021]	"
5.	"	*XhoI-HinfI (723) [1272-549]	"
6.	"	*XhoI-HinfI (1011) [1273-2283]	"
7.	"	*HinfI-HaeIII (421) [549-129]	"
8.	Yeast mitochondria	*HpaII-EcoRI (345)	S. cerevisiae (strain 5D23) Var1 gene coding region (26)
9.	" "	*EcoRI-HpaII (345)	" "
10.	pUC12	*PvuII-EcoRI (90) [306-396]	lac z gene region (27)
11.	archae-bacteria	*HpaI-EcoRI (790) [+592/-101]	5' region of methyl reductase gene of M. vannieli (28)
12.	HSV-1	*HindIII-EcoRI (352)	Ori$_s$ fragment in pON103 (29)
13.	"	*NcoI-AvaI (192) [-204/-12]	IE gene 3 promoter (30, Genbank)
14.	"	*NcoI-AvaI (121) [-204/-325]	" "
15.	"	*AvaI-NcoI (192) [-12/-204]	" "
16.	HSV-1	*AvaI-NcoI (121) [-325/-204]	as above in fragment #15

Table 1 (cont.)

17.	"	*HindIII-PvuII (228) [+11/-117]	glycoprotein D promoter region (31)
18.	X. laevis	*HindIII-RsaI (111) [560-670]	5s RNA gene 5' flank and coding regions (32)
19.	" "	*RsaI-HindIII (111) [670-570]	" "
20.	Human DNA	*AccI-SphI (245) [61816-61571]	ß-globin gene, 5' region, 52 bp purine-pyrimidine Repeat (Genbank)
21.	" "	*HpaI-AccI (444) [61372-61816]	ß-globin 5'region (Genbank)
22.	" "	*BamHI-EcoRI (918) [62666-63582]	ß-globin coding region (Genbank)
23.	" "	*EcoRI-BamHI (918) [63582-62666]	" "
24.	bacteriophage lambda	*EcoRI-BglII (359) [39168-38814]	Origin of replication (Genbank)
25.	SV40	*NcoI-RsaI (258) [37-295]	Regulatory region (Genbank)
26.	SV40	*NcoI-AvaII (162) [37-5118]	" "

[a] kDNA is kinetoplast DNA from C. fasiculata.

[b] The asterisk (*) indicates the location of the 5' end label. The size is given in base pairs. When possible, oıientation coordinates are given in brackets: + or - indicates the location of the fragment termini relative to the start site of transcription; other numbers correspond to numbered bases in the reference, for example, Genbank.

[c] kDNA sequence was provided by D. Ray and the coordinates correspond to base positions in this sequence.

sequences for each cleavage site in the data base. The analysis was carried out under a variety of constraints to reveal the frequencies and positions of palindromes. For example, looking at 9 base palindromes which allow up to 4 mismatches we found 703 in the cleavage data base (71 cleavages); however, the frequency of palindromes under these constraints or others in random sequences was not significantly different. We concluded that topoisomerase II does not require any dyad axis of symmetry in strong cleavage sites, and this is clearly reflected in the consensus sequence (however, see Fig. 3 below). We also analyzed the associative relationships among the bases flanking the cleavage site. The results of these analyses were negative: 1) Estimates of Fourier coefficients did not indicate relevant periodicities in the occurrence of nucleotides.

2) Transition probability matrices were constructed to identify Markov processes relating base identity at one position to that of another. The conditional proportions did not differ systematically from the proportions expected by chance, suggesting the absence of Markov process sequential dependencies. For example, the likelihood of a match to the consensus sequence at position n + 1 was not dependent on whether there was a match at position n (over a series of tests using the expected and observed frequencies for 71 sequences, chi-square values, with one degree of freedom, ranged from 0.0 to 2.8, with a value of 3.8 required for significance at the 0.05 level). From these data we conclude that the probabilities of base occurrence are independent of one another, and it appears that a high percentage match to our consensus sequence is what each of the cleavages has in common.

Assuming a random distribution of the four nucleotides (p = 0.25 for each), the expected frequency of matching a consensus sequence depends solely on the number of positions in the consensus sequence, the degree of base specificity at each position, and the matching criteria selected. For example, a 70% match to the consensus shown in Table 2 is predicted mathematically at one per 25.9 bases (1/probability of 70% match). Using Monte Carlo methods to count the occurrences in 196,000 bases of randomly generated data, the observed frequency of a 70% match was one per 25.8 bases. As the average match to consensus of topoisomerase II sites is 70% and the observed incidence of sites is about one per 25 bases (strong + weak sites), a consensus of the size and degeneracy shown in Table 2 is consistent with the observed frequencies of topo II sites. Similarly, the strongest topo II sites are observed at a frequency of

Table 2. Nucleotide frequencies used to derive the topoisomerase II consensus[a]

	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8	
A	.423	.225	.141	.254	.197	.141	.324	.239	.183	.056	.282	.282*	.085	.127	.310	.155	.169	.070	
C	.141	.239	.352	.169	.211	.521	.156	.113	.141	.366	.155	.113	.113	.056	.268	.282	.211	.352	
G	.268	.211	.155	.324	.169	.141	.282	.254	.521	.085	.155	.535	.451	.042	.268	.113	.211	.113	
T	.169	.324	.352	.254	.423	.197	.239	.394	.155	.493	.408	.070	.352	.775	.155	.461	.408	.465	
	A	N	C	N	N	C	N	N	G	T	N	G	G	T	N	T	N	T	
5'	G		T							C ↓			T			C		C	3'
	-10	•	•	•	•	-5	•	•	•	-1	+1	•	•	•	+5	•	•	+8	

[a] The top row of numbers are the nucleotide positions where the arrow between -1 and + 1 represents the site of cleavage by topoisomerase II. Minus positions are nucleotides 5' of cleavage and plus positions are nucleotides 3' of cleavage. The actual nucleotide proportions are given for each base listed in the column at the left margin. These data are compiled from the 71 strong cleavages in the data base. Bases taken as elements of the consensus are underscored; the asterisk indicates that at +2, G predominates, however there is an A present in more than 60% of the sites that lack a G at +2. Thus, +2 could be taken as G or A. Positions +10 and +11 (not listed) showed a preference for pyrimidines (T > C) but inclusion in the consensus lowered the average match for real data while raising the match to random data. The top strand consensus elements are shown below each position; an N is placed at each position lacking dominant nucleotides. The bottom strand is just as valid and reads: 5' A/G N A/G N ^A C/A C N A/G C N N G N N G/A N T/C 3'

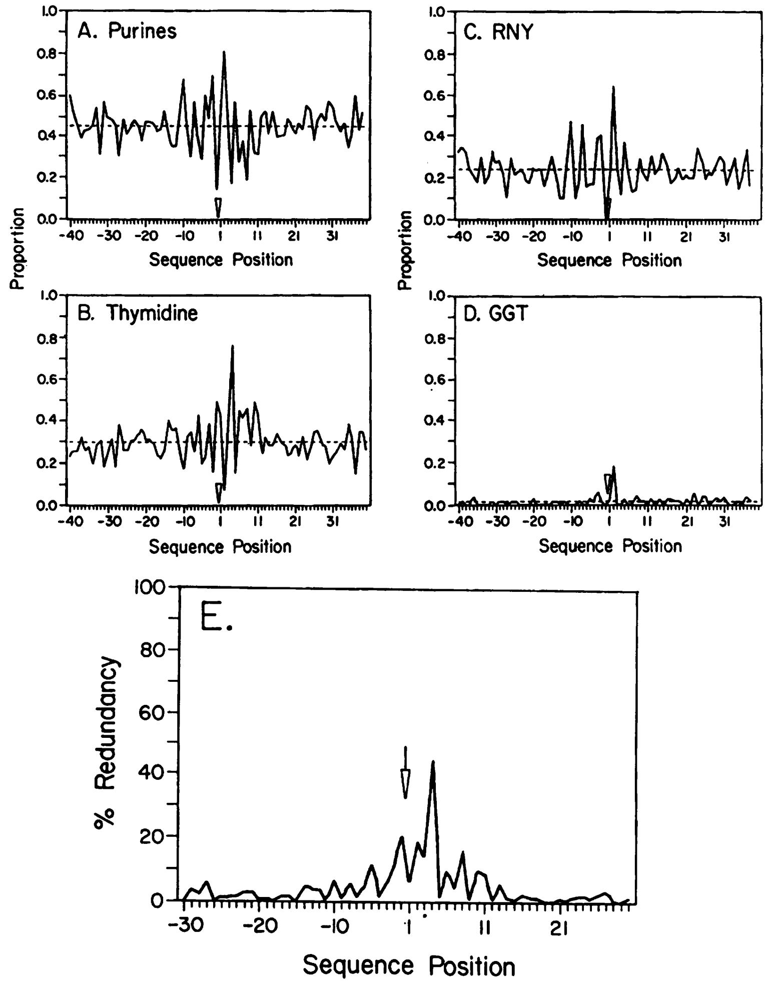

Figure 1. Analysis of nucleotide frequency and redundancy.

In figures A through D, the proportion or frequency of occurrence of purines (panel A), thymidine (panel B), purine-N-pyrimidine (panel C) and the trinucleotide GGT are shown at each position relative to the cleavage (marked on the X axis). The data were compiled from the data base of 71 strong topoisomerase II cleavages. The dotted line (- - - - -) indicates the mean proportion for the species calculated from the entire data base. Significant deviations from this mean were seen only in the consensus region. In panel E, the percentage redundancy was plotted (see "Materials and Methods" for calculations). Redundancy was plotted versus sequence position with the cleavage site located between -1 and + 1 (arrow) and the data were averaged from the 71 cleavages in the data base. The peaks show positions with redundant nucleotide proportions, i.e., positions that have a nonrandom distribution of bases. Peaks are seen only at the "non-N" consensus positions except for + 10 and + 11 (see discussion). All other positions show a redundancy near zero indicating random distributions at these positions.

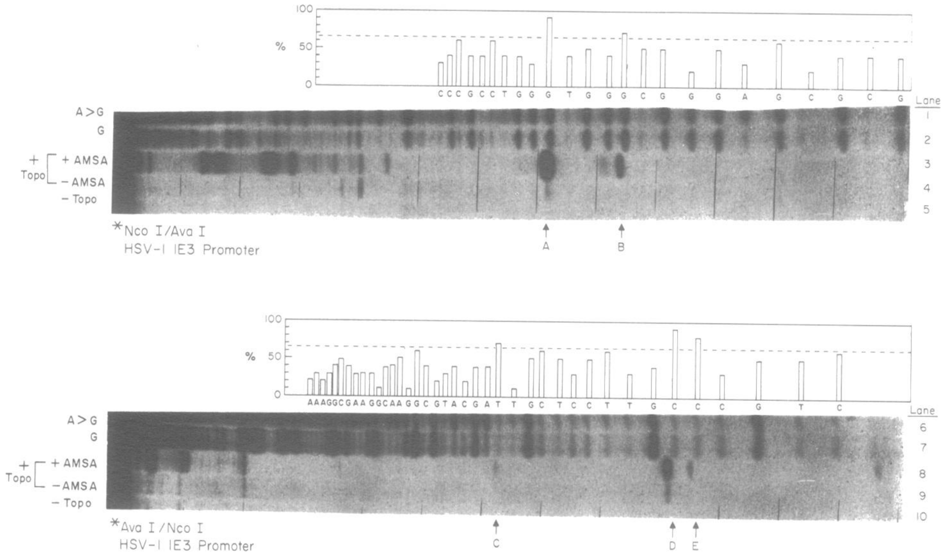

Figure 2. Topoisomerase II cleavage on HSV-1 promoter fragment.
Cleavage reactions were carried out on fragment number 14 (top gel) or fragment number 16 (lower gel) from table 1. The cleavage products were analyzed on a 12% sequencing gel. The sequencing ladders are shown in lanes 1,2,6 and 7. Lanes 3 and 8 contained 4 units of topoisomerase II plus m-AMSA; lanes 4 and 9 contained 4 units of topoisomerase II, and; lanes 5 and 10 did not receive enzyme. Above each gel is shown the actual sequence along with a histogram showing the % match to consensus at each position. Each base in the sequence was treated as +1 in the consensus (i.e., one base 3' of cleavage) with the flanking bases as -10 to +8. The consensus was moved as a window along that portion of the sequence that was resolved on gel. Only matches to "non-N" sites are included, thus a 17/18 match is actually 9/10 "non-N" positions, or 90%. The broken line (- - -) on the graphs indicates the threshold match to the consensus that appears to be required for cleavage. This value (roughly 70%) corresponds to the average match of the consensus to the real data base. The arrows designating sites A through E correspond to topoisomerase II cleavages discussed in text.

one per 130 bases and most match the consensus 80% or better, while also matching at least 40% at the site on the opposing strand; the expected odds of matching these criteria are one per 154 bases for the topoisomerase II consensus.

The consensus accurately predicts topoisomerase II cleavages

To determine how accurately the consensus sequence predicts cleavages, we selected various sequences and evaluated the homology to the consensus at each base in the test fragment. Each base was examined as a potential cleavage site so we could calculate homology to the 10 "non-N" positions; therefore, a 90% match corresponds to an actual match at 9 out of 10 specific bases even though the consensus contains 18 bases. Topoisomerase II cleavages

A

CLEAVAGE

	5' -10	-9	-8	-7	-6	-5	-4	-3	-2	-1	1	2	3	4	5	6	7	8	9	10	11	12	13	3' 14
A	.355	.277	.191	.340	.345	.211	.394	.324	.317	.106	.528	.317	.077	.268	.401	.155	.282	.155	.225	.289	.169	.331	.241	.206
C	.156	.305	.270	.199	.190	.401	.134	.162	.127	.317	.099	.282	.324	.106	.176	.401	.225	.317	.162	.176	.296	.204	.177	.284
G	.284	.177	.206	.298	.176	.162	.317	.232	.401	.176	.106	.324	.282	.099	.317	.127	.169	.134	.401	.190	.197	.275	.305	.156
T	.206	.241	.333	.163	.289	.225	.155	.282	.155	.401	.268	.077	.317	.528	.106	.317	.324	.394	.211	.345	.338	.190	.277	.355
	A	N	N	N	N	C	A	N	G	T	A	V	B	T	A	C	N	T	G	N	N	N	N	T
	G					T	G		A	C	T	= not T	= not A	A	G	T		C	A					C

B

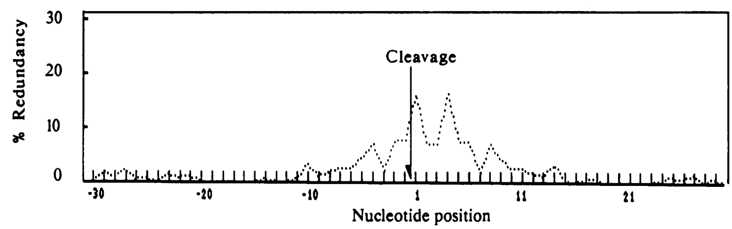

Figure 3. Analysis of nucleotide frequency and redundancy that defines a symmetric consensus sequence. The data base of 71 cleavages was analyzed for both top and bottom strands (= 142 sites) to derive a 24 base symmetric (palindrome) consensus sequence. Panel A shows the matrix of actual base proportions and below each position is an assigned consensus element. Panel B shows the redundancy analysis (see Fig. 1D, legend and Materials and Methods) which indicates that the 24 bp window (-10 to +14) embodies the key elements which most likley define the consensus sequence.

were performed and the consensus homologies at different sites compared to actual cleavage data. As shown in Figure 2, the strongest topoisomerase II cleavage band (lane 3 site labeled A) ran slightly slower than the closest guanine indicating that the cleavage was to the 5' side of the guanine. Note that because the topoisomerase II cleavage fragment has a 3' hydroxyl (8,9), its electrophoretic mobility is somewhat different than the chemical sequence marker which contains a 3'-phosphoryl end (25). The cleavage at this strong site shows a 90% match to the consensus sequence, whereas a weaker cleavage that is further 5' (lane 3, band B) has a 70% match. Other sites that match the consensus sequence in the 65-70% range were recognized and cleaved by topoisomerase II, although cleavages were usually very weak. For example, on the opposing strand (Figure 2 lanes 8,9) the strongest cleavage seen in the presence and absence of m-AMSA (band D) had a 90% match and all other sites with greater than 65% homology showed rather weak cleavages (bands C and E). Homology to the consensus does appear to reflect prominent topoisomerase II cleavages. In this particular analysis we matched only the top strand (ignoring bottom strand homology); thus, while the consensus appears to predict potential sites, the relative strength or efficiency of cleavage is not always reflected in the extent of homology to one strand and the accuracy is improved by considering both strands.

Since both strands of DNA are involved in cleavage by the topoisomerase II dimer, we analyzed the relation between match to the consensus on both strands of a cleavage site and the strength of cleavage. We catalogued the "top strand" consensus match on each of the two strands for 100 topo II sites on a variety of fragments which displayed strong, moderately strong and weak cleavages. We found that strong cleavage sites (giving intense bands in a 24 h. exposure) had an average match of 72% on the "best matched" strand; on the other strand the average match was 45%. Moderately strong cleavages (clearly visible bands after 24 h. exposure) matched an average of 58% (best matched strand) and 42% on the opposite strand, while weak cleavages (barely detectable bands in 24 h. exposure) matched 58% on one strand and 36% on the opposite strand. In addition, we analyzed all sites that showed relatively high homology to the consensus (average of 75% match) but were not cleaved by topo II and we found that these uncleaved sites were atypical in that the match to consensus on the opposite strand was very low (<30%). These findings suggest that the action of topo II is not dictated by the sequence on a single strand of DNA but rather requires that each strand be recognized (either simultaneously or independently) by one topo II subunit.

To test whether a single consensus sequence could reflect a simultaneous match to consensus on both strands of a site, we derived a symmetric consensus sequence from data on both strands of each of 71 sites in the data base to yield 142 sites. Fig. 3A shows the nucleotide frequencies at each position relative to the cleavage site from the database of 142 sequences. Fig. 3B depicts the redundancy at each position. Clearly, the symmetric data still contain significant sequence information as seen by the peaks (Fig. 3B) which were used to define consensus elements. The resulting symmetric consensus is more degenerate than the top strand (or asymmetric) consensus shown in table 2. Superficially, the symmetric consensus derived from the 142 sites resembles a blend of the top strand and the bottom strand (taking into account the 4 bp overhang); however, the latter contains much less non-random information and thus contributes more "noise". The 142 sequences matched the consensus by an average value of 74% (matching 10 to 11 of the 14 "non-N" positions), compared to an expected random match of 54%. The result is that both symmetric and asymmetric consensus sequences predicted strong cleavages equally well (9 out of 12); however, the symmetric consensus tends to predict false sites more frequently (see Fig. 4A, analysis 1).

Scoring the match of a test sequence to a consensus sequence involves a matrix of weights. In our case, the weights were either ones or zeroes because matches to single nucleotide elements in the consensus were not weighted differently than matches to multinucleotide consensus elements. This approach treats matching as an all-or-none event: Each position relative to the cleavage site is either included in the consensus model or not, and a test sequence is scored as either matching or not matching for each position included. The sum of

A

	Consensus Sequence:	Total No. of sites/100[b]	No. of Sites Accurately predicted:	No. of Sites Falsely Predicted:
Analysis 1: Based upon % match to Consensus Sequence	**Asymmetric (see table 2)**	**12**	**9**	**0**
	Symmetric (see table 3)	**12**	**9**	**2[d]**
Analysis 2: Based upon probability calculations (see results)	**Asymmetric (see table 2)**	**12**	**12**	**0**
	Symmetric (see table 3)	**12**	**10[c]**	**0**

[a] Cleavages were performed with 4 units of purified chicken topoisomerase II and fragment #6 (kDNA, table 1) either in the presence or absence of m-AMSA (data are shown in Fig. B).
[b] A total of twelve cleavages were selected for this analysis and are marked below in Fig. B.
[c] The sites which we failed to predict were relatively weak sites; for example in this case, sites 8 and 9 (Fig. B) were not predicted according to our threshold conditions (see text).
[d] These two sites were predicted by a favorable match to consensus, but were not detected in the experiment in Fig. B.

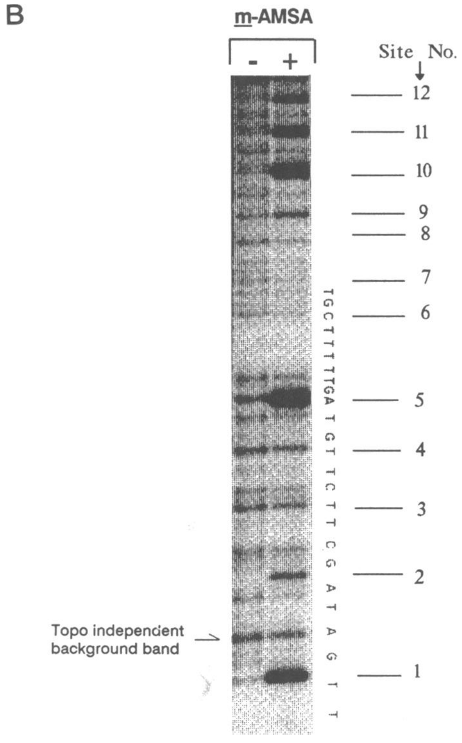

Figure 4. Comparison of Symmetric and Asymmetric Topo II Consensus Sequences to Actual Sites A Site predictions[a] B Sequenced Topo II sites in kDNA:

Table 3. Comparison of the Symmetric and Asymmetric Consensus Sequences to the Data Base.

Consensus Sequence:	Total Number of sites analyzed:	Actual number of Topo II Sites:	Number of Sites accurately predicted	Number of Sites falsely predicted
Comparison Based on % match to consensus sequence				
[a]Asymmetric Consensus Sequence:	[c]1200	[d]66 [14]	33 [13]	19
[b]Symmetric Consensus Sequence:	1200	66 [14]	32 [13]	46
Comparison Based on Probability Calculations				
Asymmetric Consensus Sequence:	1200	66 [14]	29 [13]	15
Symmetric Consensus Sequence:	1200	66 [14]	26 [13]	19

[a]Asymmetric Consensus Sequence: 5' R N Y N N C N N G Y ^ N G $^{G}/_{T}$ T N Y N Y 3'
[b]Symmetric Consensus Sequence: 5' R N N N N Y R N R Y ^ $^{A}/_{T}$ V B $^{T}/_{A}$ R Y N Y R N N N N Y 3'
(V = not T; B = not A)

[c]This value corresponds to the total number of nucleotides in the data base used in this comparison.
[d]The cleavages were tabulated from sequencing gels of cleavage reactions that were selected at random from the data base of eukaryotic sequences. This value corresponds to strong and moderately strong cleavages. Strong cleavages are given in the brackets.

the matching positions is then expressed as a percentage of the total number of positions included in the consensus sequence. In any given analysis, the actual number of sites predicted depends upon the thresholds chosen and we selected the requirements for match as follows: For the asymmetric sequence, ≥11 matches after combining independently derived matches to the top and bottom (for example, 7/10 match to top and 4/10 match to bottom strand); for the symmetric consensus sequence, the threshold was 10 out of 14 "non-N" positions (only one strand is required since the other match is identical). These thresholds were selected after many trials because they predict all strong sites and about half of the weak sites while excluding the most false sites.

An inherent feature of any consensus sequence is that it does not represent all of the information that can be derived from the data. More of the information in the composite sample of cleaved sequences may be retained by using a matrix based on the observed frequencies of nucleotides at each position relative to the cleavage site (47-51). We used this approach by determining which N locations had nonrandom base frequencies and then constructing a 4 x N matrix of the logarithms of the proportions of each nucleotide at each position (see Table 2 and Figure 3A). With this approach, the degree of match of a test sequence is computed by summing the matrix values corresponding to the test sequence bases at each position included in the model. The antilogarithm of the sum of these values then equals the probability of obtaining the particular test sequence based on the assumption that the test sequence is a topo II cleavage site. For example, if in a test sequence, A at -1 and T at +1 is encountered, then values of 0.106 and 0.268 are assigned respectively (see Figure 3A matrix). The assigned values at each position in the sequence being analyzed can be summed directly, but the probability model (in which the natural logarithms are summed) is more valid for prediction of topo II sites. Probability calculations were performed by scoring the top strand and the corresponding cleavage on the bottom strand and recording an average of the two. If the scores for the potential

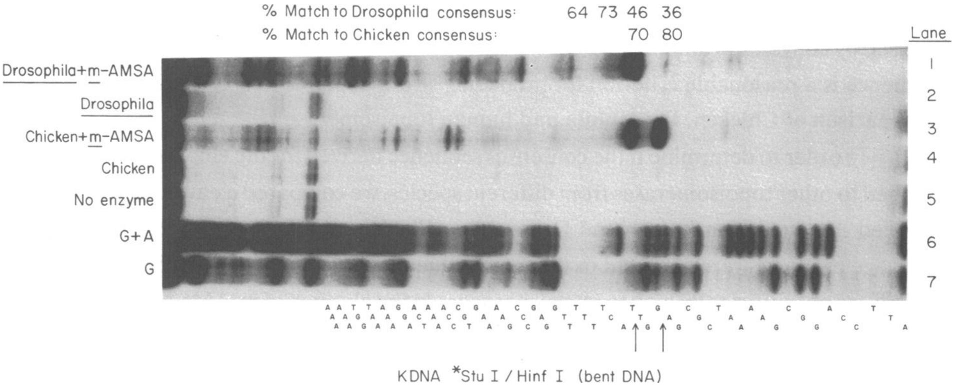
% Match to Drosophila consensus: 64 73 46 36
% Match to Chicken consensus: 70 80
Lane
Drosophila+m-AMSA
Drosophila
Chicken+m-AMSA
Chicken
No enzyme
G+A
G
1
2
3
4
5
6
7
A A T T A G A A A C G A C G G T T T G C T A C A T
A A G A A G C A C G A A C A T T C T A G A A G C T
A A G A A A T A C T A G C G T T A G G C A G G C A
KDNA *Stu I / Hinf I (bent DNA)

Figure 5. Comparison of chicken and Drosophila topoisomerase II cleavages. Cleavages were performed on fragment number 1 (table 1) using approximately equal amounts of topoisomerase II activity (2 units) of the two enzymes. The gel (run from left to right) contains the following reactions: Lanes 1, 3 and 5 contained either Drosophila, chicken or no enzyme, respectively and m-AMSA; lanes 2 and 4 show Drosophila or chicken enzyme cleavages, respectively, without drugs. Sequence ladders are marked in lanes 6 and 7. Actual sequence is shown below the gel. There are a few higher molecular weight fragments (lane 5) that are due to degradation of the parent DNA fragment; therefore, cleavages were not evaluated in this size range.

sites on a sequence are rank-ordered from largest to smallest, the larger the value the greater the likelihood that the site is an actual topo II cleavage site, particularly for the very largest values (which are often strong cleavage sites). The probability method is more valid for prediction of sites if minimum threshold values are used.

The cleavage data in Fig. 4B were analyzed by the probability model using the antilogarithm threshold values of 3 x 10^{-6} for the asymmetric consensus matrix and 1 x 10^{-8} for the symmetric consensus matrix. The results (Fig. 4A, analysis 2) reveal that both symmetric and asymmetric matrices predict sites, although the latter is more reliable. We compared the reliabilities of different prediction models for a data base of 1200 bp containing 66 topoisomerase II sites (14 were strong sites) that were not in the original data base. As shown in Table 3, the number of accurately predicted sites is similar for all 4 methods (using optimized thresholds described above) in that all predict about 50% of the total sites (strong plus weak) and 13/14 strong sites; however, the false rates are variable. Parenthetically, the selected threshold values affect reliability. For example, if the acceptable probability for the asymmetric matrix is lowered to 1 x 10^{-6}, then 39 of the total 66 sites are predicted as well as all 14 strong sites, but the false site value increases from 15 to 39; in contrast, if the value is raised to 1 x 10^{-5}, 16 out of 66 total sites are identified with only 3 false sites. In sum, the percentage consensus match and the matrix probability model are both reliable methods for predicting sites with the asymmetric consensus data. Since the use of consensus sequence data is so prevalent, it is likely that this method will be used by convention and we have shown that percentage match to an asymmetric consensus sequence is a reasonable criterion for predicting strong topoisomerase II sites.

Comparison of Chicken, Drosophila and human topoisomerase II

In order to determine if the consensus sequence derived for the chicken enzyme can be applied to other topoisomerases from different species, we compared cleavage products using purified enzymes from human and Drosophila. Chicken and Drosophila topoisomerase II cleavage products were analyzed in the same sequencing gel (Figure 5). Enzyme concentrations were not sufficiently high to reveal cleavages in the absence of m-AMSA. It appears that cleavage sites are very similar between the two topoisomerases. The chicken consensus again accurately predicted cleavages by the chicken enzyme. The two strongest sites (Figure 5, lane 3) showed a 70 and 80% homology to the chicken consensus. The strongest Drosophila topoisomerase II site (lane 1) matched its published consensus (as determined in the absence of drugs [42]) at 46% of "non-N" sites. In addition, the Drosophila consensus showed a 73 and 64% match at sites 3 and 4 bases, respectively, 3' of the two cleavages; however, we did not detect any cleavages at these sites. It is clear that there are exceptions to the consensus for both the chicken (see Figure 4B) and Drosophila data (Figure 5). Nonetheless, these data show that the two enzymes display a fairly well conserved set of binding sites although some differences in specificity and cleavage efficiency can be demonstrated.

A comparison of chicken, Drosophila and human topoisomerase II cleavages produced on several DNA fragments is shown in Figure 6. In all cases the human and chicken enzymes cleaved the same sites and with the same relative strengths, while the Drosophila topoisomerase II cut at somewhat different sites with quite different strengths. This suggests that topoisomerase II sites are partially conserved among highly divergent species and essentially identical among vertebrates. Thus, the chicken topoisomerase II consensus sequence is a better predictor of cleavage sites by a vertebrate compared to an invertebrate enzyme. The consensus is likely to apply in vivo as well, because in vivo sites have been reported to be a subset of the in vitro sites (12,15).

The ability of our consensus to predict cleavages in vivo is suggested by mapping experiments of topoisomerase II cleavages on the SV40 minichromosome in situ (12). Endogenous topoisomerase II cleaves at a major site around nucleotide 270 in the SV40 control region (12; M. Gallo, M. Muller and J. Spitzner, unpublished data). A homology search of this region of SV40 revealed a 90% match to our consensus sequence with a cleavage between nucleotides 273 and 274 on the early strand. Strong cleavage at this site can be demonstrated in vitro (data not shown). Additional sites within the SV40 genome have been identified which show a 90% match to our topoisomerase II consensus. It is striking that out of 10 sites identified in SV40, 6 are clustered within the control region or in closely flanked sequences. Within the

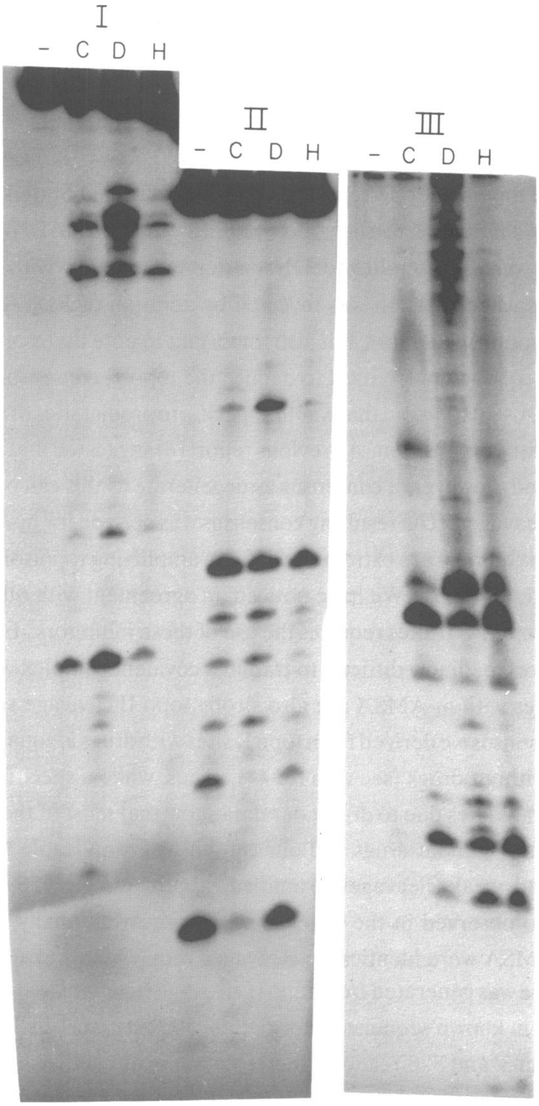

Figure 6. Comparison of topoisomerase II from different organisms.
Different labeled fragments were tested with either no enzyme (dash), chicken enzyme (C), Drosophila enzyme (D), or human enzyme (H). Gels I, II and III contain fragment #13, #14 and #1 respectively, from table 1. Cleavage products were resolved on a 6% sequencing gei.

control region, topoisomerase II consensus sites (90% matches) are present 5' of the replication origin at nucleotide 5051, in the 21 bp repeats at nucleotide 64 and two sites 3' of the 72 bp repeats at positions 273 and 407.

DISCUSSION

Derivation of the topoisomerase II consensus.

Topoisomerase II cleavages were catalogued on a wide range of DNA substrates. The following observations indicated that these cleavages were generated by topoisomerase II: The enzyme used was purified to homogeneity ($M_r = 155,00$) and was shown to alter DNA linking number of a unique topoisomer in steps of two (Muller et al., submitted); the cleavages were enhanced by drugs that stabilize the topoisomerase II/broken DNA intermediate; the cleavage products were protein-linked DNA after termination with SDS (Muller et al., submitted); the cleavage reaction was reversed by addition of EDTA or high salt prior to addition of SDS (unpublished data). It is also important to note that a relatively large number of sites of diverse sequences were used to deduce the topo II consensus. If fewer sites were evaluated, a consensus was derived that was too specific to predict sites of cleavage in sequences of vastly different base composition. A previous report (8) suggested that a consensus sequence could not be deduced; however, a consensus as degenerate as this can only be derived from a larger number of cleavages. The resulting consensus was derived for m-AMSA induced cleavages because these inhibitors are extremely useful in amplifying topoisomerase II sites both in vivo and in vitro (11,13,15,33). We have noticed, in agreement with others (15), that in vivo mapping of topoisomerase II sites requires the use of these inhibitors. Evidently, endogenous topoisomerase II is exceedingly difficult to trap in a covalent complex without the inhibitors. The strong cleavages with m-AMSA are also strong topo II cleavage sites in the absence of inhibitors. The consensus we derived from topo II sites with drugs is equally valid for predicting topo II cleavages without drugs (see also ref 44). Thus, while a specific sequence may show differences in cleavage sites due to drugs, our data show that most of the strong cleavage sites are identical with and without drugs and our consensus is applicable to both. (Note, in 230 different strong and weak cleavages examined, approximately 75% of all drug-induced cleavages were also observed in the absence of drugs. Additionally, 85% of all cleavages detected with m-AMSA were identical to cleavages with VM26 and vice versa.)

The data base was generated from topo II cleavage reactions as described in "Materials and Methods". From known sequence information, 50 bases were included on either side of the cleavage point in 3' and 5' directions, therefore, we evaluated a large amount of flanking sequence to ensure that essential elements of recognition were not overlooked. The following pieces of evidence indicate that the 18 bp of sequence are sufficient for consensus definition. First, cleavages were detected within 10 bp of fragment ends. Second, a cleavage site in a 30 bp synthetic oligonucleotide was no different when the same site was embedded in a larger fragment. Third, cloned fragments with cleavages near ligation sites in the vector showed the same site specificity before and after cloning. These observations are consistent with the analysis of redundancy of bases that flank the cleavage site (Figure 1E) .

Topoisomerase II makes transient double strand breaks in DNA; however, it is not known to what extent homology on each strand is important in site definition. Recently obtained data show that topoisomerase II can be trapped in a covalent complex with single strand nicks at sites which fit the consensus sequence (Muller et al., manuscript submitted); therefore, consensus elements may independently contribute to recognition on each strand. Indeed, the analysis of match to consensus on both strands of topo II sites indicates that contributions from both strands affect the sites cleaved and the efficiency of cleavage. Upon inspection of both strands in the data base, the following patterns were noted: Sequences showing the highest homology match on one strand (average of 7 of 10 consensus positions) and a moderate match on the other strand (average 4 or 5 out of 10) were the strongest cleavage sites; sequences with less homology on one strand (average 6 out of 10) and a moderate match on opposite strand (average 4 of 10) were cleaved with moderate efficiency; sequences with an average of 6 out of 10 matches on one strand but rather low opposite strand homology (average 3 to 4 matches) were associated with very weak cleavage sites. Finally, sequences with high matches (7 to 8 of 10 consensus elements) on one strand but only three or fewer matches on the opposite strand were generally not cleaved by topo II, while on average, uncleaved sites matched only 4 out of 10 as the best match with 3 out of 10 on the opposite strand. Thus, we have seen evidence for additive effects of consensus elements on opposing strands.

Further examination of both cleaved strands of topo II sites revealed additional details about individual consensus elements. In 70 (of 71) strong cleavages in the data base, a pyrimidine was found at position -1 on at least one strand. Also, 67 of the 71 sites had a pyrimidine at +4 (or purine at +1 on opposite strand) while the other 4 sites contained a pyrimidine at +3. All 71 had at least one of the +6 or +8 pyrimidines on one strand or the other and there were 8 sites that lacked a +6 pyrimidine but these had pyrimidines at +7 as well as +8. An additional 4 sites without pyrimidines at +8 had them at +2, +3 and +6. Finally, 65 of 71 had a purine at +2, and the exceptions (6 sites) had a purine at -2. The sample size is perhaps not large enough to determine the significance of all the relations noted above, but these observations, taken together with the correlation between strength of cleavage and match to consensus on both strands, imply that each monomer of topo II recognizes some of the same positions on each strand of DNA during cleavage. This led us to examine topoisomerase II cleavages as double stranded events involving consensus elements on both strands. We used nucleotide proportions for both strands of the 71 strong cleavages (142 sites) to derive a symmetric consensus sequence which minimizes bias introduced by selecting one single strand as the top strand (see results). The symmetric consensus sequence did not predict topoisomerase II sites as accurately as the asymmetric consensus sequence (see Table 3 and Fig. 4) which separately examines the two strands of the helix for homology. We found that using the matrix of base proportions, which makes use of all the information in the data base, was about as reliable as consensus match so long as the asymmetric data were used and the logarithms of frequencies were added only for positions with non-random nucleotide frequencies (note, a variety of consensus matrices were also tested, but utilizing data at more positions decreased the validity of the model). The asymmetric single strand consensus model seems preferable for prediction because of its simplicity and the ease of conveying information in terms of a consensus. As a first estimate, a preliminary search for matches of 80% or better will provide valuable information about potential topoisomerase II sites. For example, a search for 80% matches to the consensus sequence frequently will detect stretches of alternating purine-pyrimidine sequence (note that the consensus sequence will fit alternating purine/pyrimidine DNA). As a test of the significance of these findings, we recently demonstrated that topoisomerase II is acutely reactive toward sequences that contain poly purine/pyrimidine regions (Muller, Spitzner and Chung, ms. in preparation).

Comparison of cleavage specificities and consensus homology with vertebrate and invertebrate topoisomerases

Drosophila and chicken topo II cut similar sites with m-AMSA but with quite different efficiencies. In a direct comparison (Figure 5), the strongest m-AMSA chicken topo II cleavage matched the consensus at 80% and was visible (although faintly) in the absence of drug. The same site was recognized and weakly cleaved by the Drosophila topo II. A second chicken enzyme site (70% match) represents the only other match in this fragment and this site was a strong Drosophila site but matched the Drosophila consensus at 46%. The experimental data suggest that although the two enzymes may have different consensus recognition sequences, they are not highly divergent. On the other hand, we cannot rule out in vitro artifacts (proteolytic cleavages) or post-translational modifications of the enzyme (34) which might make the two enzymes behave differently. In contrast, the human and chicken enzymes sites are the same (see Fig. 6); therefore, the consensus is probably valid for vertebrate species.

The topoisomerase consensus and enhancers

Studies on mapping endogenous topoisomerase II cleavages have revealed that in vivo sites are most likely a subset of in vitro sites (12,13,15; M. Muller and V. Mehta manuscript in preparation). The basis for site selection in vivo (besides sequence) is unknown but probably related to site accessibility, e.g., nucleosomes or non-histone chromosomal proteins may occlude potential sites. Site distribution is non-random with respect to protein coding and non-coding regions (14) and sites are enriched in intergenic regions. Clustered in vitro cleavages by topoisomerase II were also reported in an enhancer region of c-fos (33) and the calf thymus topoisomerase II cleavages showed an average of a 60-65% match to our consensus sequence. Additional confirmation comes from mapping endogenous topo II sites where cleavages frequently (but not always) align near nuclease hypersensitive regions (13,15) and in SV40 where the strongest in vivo site (near nucleotide 270) is present in the control region (12). We have confirmed the in vivo result (data not shown) and a consensus search of the region revealed a 90% match at position 273, which is within domain B of the enhancer (35). These correlations

Table 4. Frequency of topoisomerase II consensus in enhancers

Enhancer:	Real Sequence[a]		Random Sequence[b]	
	100%	90%	100%	90%
Adenovirus Ela (126 bp) (45)	0	1	0	0
RSV LTR (296 bp) (Genbank)	0	1	0	0
Human Immuno- globulin (319 bp) (Genbank)	0	2	0	0
Mouse Immunoglobulin (kappa) (479 bp) (43)	1	1	0	2
SV40 (strain 776) (183 bp) (39)	0	1*+	0	0
Chicken ß^A-globin 3' enhancer (484 bp) (38)	0	1*	0	0
HSV-1 IE gene 3 (230 bp) (40)	0	3+	0	0
**Human ß-globin 3' (700 bp) (Genbank,41)	0	0	0	0

[a] Matches of consensus to the actual sequence of enhancer.

[b] Matches of consensus to randomized sequence of identical base proportions.

*In vivo site mapped.

+ In vitro site confirmed by cleavage assays.

**90% Match identified 159 bp upstream of enhancer sequence.

are striking for several reasons. First, 6 out of 10 sites showing 90% homology to our consensus, are located in an 800 bp domain centered over the control region of SV40. Second, the 90% match in domain B of the SV40 enhancer (35,36) is also the major site in vivo in minichromosomes (12). Third, work from Chambon's lab has identified a number of trans-acting factors that recognize the region GT-II that encompasses the strong topo II cleavage homology (37). The factor GT-IIA exists in different cell lines and methylation interference experiments revealed key G residues at positions 272, 275 and 276. These sites are aligned below with our consensus sequence (G residues that interfere with GT-IIA factor binding are underscored):

base:	270	271	272	273	274	275	276	277	278	279	280
SV40:	C	A	G	C	T	G	G	T	T	C	T
Topo:	N	N	G	Y	N	G	G/T	T	N	Y	N

Note that methylation interference is seen at conserved G residues that match the topoisomerase II consensus. Furthermore, the GT-IIA protein/DNA complexes cannot be effectively competed by fragments that do not match the consensus sequence but are competed by fragments with homology to the consensus sequence.

These observations prompted us to survey enhancer regions in various DNA sequence data bases for topoisomerase consensus sequences. In most enhancer sequences we found one or more sites of 90% or greater homology. To evaluate this further, we selected several well characterized enhancers (Table 4) and determined the frequency of 90% and 100% matches to topo II sites. As a control, we had the computer generate random sequences of identical base proportions for each enhancer. We found only two 90% matches in the randomized data and no 100% matches, whereas the enhancer sequences contained 11 matches (1 at 100% and 10 at 90%). Assuming that this is a representative sample of viral and cellular enhancers, it appears that the topo II consensus is enriched in these cis-active sites. It is particularly significant that in the SV40 enhancer (12) and the chicken β^{A}-globin enhancer (Muller, M. and Mehta, V., manuscript in preparation) the 90% matches correspond to sites that are recognized by topo II in vivo. The prevalence of topoisomerase II sites (based upon consensus homology) in enhancer regions is not seen in other sequences. Notably, in 70 KB of contiguous human DNA on chromosome 11, on the average we find 90% matches to the consensus every 700 bp, although the distribution is somewhat irregular and there are in some cases very long gaps between matches.

Present knowledge suggests an association between topoisomerase II consensus sites and enhancers; however, the reason for this association remains to be elucidated. The two may not be functionally related. For instance topoisomerase II is enriched in the nuclear scaffold (21-23) perhaps because it must act on "scaffold associated regions" (SARs) as a means of decatenating daughter chromatids. If enhancer regions are also localized near SARs, then topoisomerase II may not play a role in enhancer/promoter function. As an alternative, it is possible that locating topoisomerase II in SARs places the enzyme (and associated sites) at key positions to adjust DNA topology within domains. It is notable that in two instances where in vivo mapping has been performed (SV40 and chicken β^{A}-globin), the in vivo sites are embedded in sequences where there exists a number of potential topo II sites (predicted by consensus) but, in both cases, endogenous topo II cleaves only in the enhancer.

MATERIALS AND METHODS

Materials Restriction enzymes came from Bethesda Research Laboratories; T4 polynucleotide kinase was from United States Biochemical Corporation; reagents for Maxam and Gilbert sequencing (piperidine and dimethylsulfate) were from Aldrich; [gamma-^{32}P]-ATP came from ICN; BioRex 70 and polyacrylamide were from BioRad Laboratories; and Phenyl Sepharose was from Sigma.

Enzyme Purification The procedure for purification of chicken topo II is described in detail elsewhere (Muller et al., submitted). A similar procedure was used to purify human placenta topo II as follows. A single human placenta was minced with scissors and washed in sterile water to remove erythrocytes. Chunks of tissue were blended (Waring blender, two 30 sec bursts at maximum speed) in cold buffer [10 mM Tris-Cl (pH 7.6), 1 mM EDTA, 5 mM $MgCl_2$, 1 mM phenylmethylsulfonyl fluoride (PMSF)]. The extract was filtered through several layers of cheese cloth and then nuclei and some debris were pelleted by centrifugation (6000 xg) and resuspended in 50 mM Tris-Cl (pH 7.6), 50 mM $NaHSO_3$, 1% ß-mercaptoethanol, 1 mM PMSF. Solid NaCl was then added to 0.4 M and the suspension stirred 60 min on ice followed by addition of polyethyleneimine (0.1% final) and after 20 min of vigorous stirring on ice, the suspension was centrifuged 20 min at 10,000 xg. Ammonium sulfate was added slowly to the supernatant until 70% saturation and after stirring for 45 min on ice, the mixture was centrifuged 20 min at 10,000 xg and the pellet resuspended in buffer A [0.1 M potassium phosphate (pH 7.1), 10% (v/v) glycerol, 25 mM ß-mercaptoethanol, 10 mM $NaHSO_3$, 0.5 mM PMSF]. The extract was then loaded onto a 50 ml BioRex 70 column that had previously been equilibrated with 0.2 M buffer A (buffer A containing 0.2 M phosphate buffer). After washing with equilibration buffer, the activity was eluted in a single step with 0.6 M buffer A. This eluate was then loaded onto a 4 ml Phenyl Sepharose column equilibrated with buffer B [20% (v/v) ethylene glycol, 25 mM potassium phosphate (pH 7.1) 10 mM $NaHSO_3$, 25 mM ß-mercaptoethanol, 0.5 mM PMSF] then washed 10 column volumes of buffer B. Topo II was desorbed with buffer B containing 60% ethylene glycol. Topoisomerase II activity was diluted to 1 unit per ul where one unit is defined as the amount that will completely decatenate 0.25 ug of kinetoplast DNA in 30 min at 30° C. Analysis of polypeptides by SDS-PAGE revealed a predominant band of M_r = 160,000 and several minor bands of considerably lower molecular weight. Drosophila topoisomerase II was a gift from N. Osheroff.

Cleavage Reactions Topoisomerase II cleavage reactions were performed in a final volume of 20 ul in a standard cleavage buffer containing the following solutes: 30 mM Tris-Cl (pH 7.6), 60 mM KCl, 15 mM ß-mercaptoethanol, 8 mM $MgCl_2$, 3 mM ATP, 30 ug BSA/ml. DNA substrates were 5' end labeled with [gamma-^{32}P]-ATP and polynucleotide kinase using standard protocols and cleavage reactions contained between 2.5 to 10 x 10^4 DPM (ca. 2 to 8 ng DNA). The reactions were assembled on ice and initiated by addition of 4 units of purified topo II and where indicated immediately followed by addition of 1 ul of stock solutions of the following topo II inhibitors: Epipodophyllotoxin, VM26 (from the National Cancer Institute) prepared as a stock solution at 10 mg/ml or m-AMSA [4'-(9-acridinylamino)methanesulfon-m-anisidide] which was prepared as a stock solution in DMSO at 1 mg/ml. Reactions without drugs run with 1 ul of solvent (DMSO). Reactions were then immediately placed at 30° C and after 3 to 30 min (incubation periods over this range gave identical results) reactions were terminated by addition of 40 ul 1.5% (v/v) sodium dodecyl sulfate followed by digestion with 50 ug proteinase K/ml. Samples were incubated 30 min at 65° C, followed by addition of 0.1 vol sodium acetate (3 M), 0.1 volume of 0.1 M $MgCl_2$ and 2 vol of 95% ethanol. Samples were placed at -70° C for 15 min and centrifuged at 13,000 xg for 15 min. The DNA pellets were dried and resuspended in a sequencing gel loading buffer (25) at 1 x 10^4 DPM/ul, then boiled for 3 min and ice quenched. The samples were loaded onto a sequencing gel of appropriate porosity for the size fragment being tested and each gel contained chemical sequence latters as markers.

Statistical Methods and Analysis of Data Cleavage sites were resolved to the base by comparison of cleavage bands to sequencing markers: The cleavage event is on the 5' side of the sequencing base which runs just faster than the cleavage band because it contains a terminal 3' phosphate. This base was chosen as "+1" which places the cleavage site between +1 and -1. The strongest cleavages (the most intense bands that can be seen after an overnight exposure) were analyzed and the sequence information catalogued with up to 50 bases 5' and 3' of the site stored in a data base. Sequences were aligned so that the cleavage was between positions 50 and 51 and in those cases where the cleavage was near a fragment end, "N"s were inserted as fillers. This approach was used to compile a data base of 71 strong cleavage sites in a variety of different DNA fragments (see Table 1).

The data base was aligned to "top strand" cleavages by an iterative process (see results section). The end result of this process was a Table of nucleotide frequencies. At each individual position (such as -2, which is position 49 in the data base) nucleotide frequencies were averaged for the 71 sequences. From these data, positions with a single nucleotide proportion greater than 50% or with 2 nucleotides combined greater than 70% were scored as consensus elements as described in results. To determine whether the arbitrarily chosen values of 50% and 70% were in fact meaningful, the percentage match to the consensus was calculated for each cleavage in the data base. This was done by comparing the base at each position of a cleaved sequence with the base at same position in the consensus sequence. These data were tabulated for the all cleavages in the data base. Matches were scored without weighting, so that a match to a pyrimidine at -1 counted the same as a match to C at -5. Matches to N's were not counted, so that matches to the consensus correspond to the number of matches to the 10 "non-N" positions in the consensus. Thus, a 90% match corresponds to 17 out of 18 total sites or 9 out of 10 "non-N" sites. Matches to random sequence data were done in the same fashion, with a computer generated random data base containing the same overall nucleotide proportions as the real data base.

The 71 cleaved sequences matched the topo II consensus in Table 2 with an average value of 70%; random sequences match this consensus by an average of 43%. Because the criteria for selecting the consensus sequence were somewhat arbitrary, many permutations of the Table 2 consensus were evaluated. This consensus represents the largest sequence with the highest ratio of real data match to random data match. Some consensus positions are not intuitively obvious. For example, although pyrimidines at positions 10 and 11 could have been included (based upon frequency), doing so decreased the average match to the real data and increases the match to random sequence data. Similarly, making the individual consensus elements more specific (for example by raising the frequency requirements for dinucleotides from 70% to 80% or by requiring a T at -1 instead of T/C), resulted in a lower percentage match to both real data and random data. In these cases, the consensus was not as effective a predictor of topo II cleavages.

An analysis of redundancy was carried out to ascertain the number of flanking nucleotides 5' and 3' of the cleavage site that should be considered for generating a consensus (46). Redundancy analysis condenses the information content at a position (for the entire data base) into a single expression which describes the consistency of base occurrence. Redundancy is derived mathematically in the following way.

(1) $U(x) = [-p(x_i)\log_2 p(x_i)]$

Where U(x) is uncertainty, $p(x_i)$ is the probability of a specific base occurring at a particular position and x has four possible out comes (i= 1,2,3 or 4) corresponding to the four bases. As there are four possible outcomes, the maximum uncertainty at a given position is (from equation

(2) $(4)(-0.25)(-2) = 2.0$

The percent redundancy therefore is: $1 - \frac{\text{Actual U(x)}}{\text{Maximum U(x)}} \times 100\%$

A redundancy near zero indicates a random nucleotide distribution at a position, while a greater redundancy indicates some degree of consistency at a position; this information is graphed in Figure 1A. If nucleotide proportions at a position are essentially random, then redundancy approaches zero (see Figure 1). All of the analytical methods discussed were conducted using Pascal implementations of the relevant algorithms on IBM PC/XT computers.

ACKNOWLEDGEMENTS

The work was supported by Public Health Service grant GM31640 from the National Institutes of Health by grant #1-1031 from the March of Dimes Foundation. The authors are grateful to the following individuals who provided reagents: Neil Osheroff (purified Drosophila topoisomerase II); P.S. Perlman (yeast mitochondrial clones and sequence data); D. Ray (kinetoplast sequence data); D. Brown (5s rRNA gene); E. Vanin (human globin gene region); J. Reeve (archaebacterial genes and sequence information); C. Preston and D. McGeoch (HSV-1 immediate early genes and sequence data); and E. Mocarski for pON 103. Helpful discussions of P.S. Perlman, B. Sugden are also acknowledged. The authors are indebted to J.H. Spitzner of TEAM associates (Westerville, OH) for software implementation of the methods used in the statistical analysis of the data and for critical review of the manuscript. The authors acknowledge the helpful advice of G. Stormo concerning the data analysis.

*To whom correspondence should be addressed

[a]**Abbreviations**: Topoisomerase II, topo II; SDS, sodium dodecyl sulfate; SDS-K+, precipitate produced upon mixing of KCl and SDS; m-AMSA, 4'-(9-acridinylamino)methanesulfon-m-anisidide.

References

1. Miller, K.G., Liu, L.F. and Englund, P.T. (1981) J. Biol. Chem. **256**, 9334-9339
2. Goto, T. and Wang, J.C. (1982) J. Biol. Chem. **257**, 5865-5872
3. Shelton, E.R. Osheroff, N. and Brutlag, D.L. (1983) J. Biol. Chem. **258**, 9530-9535.
4. Benedetti, P., Baldi, M.I., Mataoccia, E. and Tocchini-Valentini, G.P. (1983) EMBO J. **2**, 1303-1308
5. Duoc-Rasy, S., Kayser, A., Riou, J.-F. and Riou, G. (1986) Proc. Natl. Acad. Sci. USA **83**, 7152-7156
6. Vosberg, H. P. (1985) in Current Topics in Microbiology and Immunology Vol. 114, pp. 19-101, Springer Verlag, NY
7. Wang, J.C. (1985) Annu. Rev. Biochem. **54**, 669-697
8. Liu, L.F., T.C. Rowe, L. Yang, K.M. Tewey, and G.L. Chen. (1983) J. Biol. Chem. **258**, 15365-15370
9. Sander, M. and Hsieh, T. (1983) J. Biol. Chem. **258**, 8421-8428
10. Chen, G.L., L. Yang, T.C. Rowe, B.D. Halligan, K.M. Tewey, and L.F. Liu. (1984) J. Biol. Chem. **259**:13560-13566
11. Nelson, E.M., Tewey, K.M. and Liu, L.F. (1984) Proc. Natl. Acad. Sci. USA **81**, 1361-1365
12. Yang, L., T.C. Rowe, E.M. Nelson, and L.F. Liu. (1985) Cell **41**:127-132
13. Rowe, T.C., J.C. Wang and L.F. Liu. (1986) Mol. Cell. Biol. **6**, 985-992
14. Udvardy, A., Shedl, P., Sander, M., and Hsieh, T. (1985) Cell **40**, 933-941
15. Udvardy, A., Schedl, P., Sander M., and Hsieh, T. (1986) J. Mol. Biol. **191**: 231-236
16. Uemura, T. and Yanagida, M. (1984) EMBO J. **3**, 1737-1744
17. DiNardo, S., Voelkel, K. and Sternglanz, R. (1984) Proc. Natl. Acad. Sci. USA **81**, 2616-2620
18. Holm, C., Goto, T., Wang, J.C., and Botstein, D. (1985) Cell **41**, 553-563
19. Duguet M., Lavenot, C. Harper, F., Mirambeau, G. and De Recondo, A. (1983) Nucleic Acids Res. **11**, 1059-1075
20. Uemura, T. and Yanagida, M. (1986) EMBO J. **5**, 1003-1010
21. Earnshaw, W.C. and Heck, M.S. (1985) J. Cell. Biol. **100**, 1716-1725
22. Berrios, M., Osheroff, N. and Fisher, P.A. (1985) Proc. Natl. Acad. Sci. USA **82**, 4142-4146
23. Gasser, S.M., Laroche, T., Falquet, J. Tour, E.B., and Laemmli, U.K. (1986) J. Mol. Biol. **188**, 613-629
24. Cockerill, P.N. and Garrard, W.T. (1986) Cell **44**, 273-282
25. Maxam, A.M. and Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA **74**, 564-569
26. Hudspeth, M., Vincent, R., Perlman, P., Shumard, D. Treisman, L. and Grossman, L. (1984) Proc. Natl. Acad. Sci. USA **81**, 3148-3152
27. Yanisch-Perron, C., Vieira, J. and Messing, J. (1985) Gene **33**, 103-119
28. Cram, D.S., Sherf, B.A., Libby, R.T., Mattaliano, R.J., Ramachandron, K.L. and Reeve, J.N. (1987) Proc. Natl. Acad. Sci. USA **84**, 3992-3996
29. Elias, P., O'Donnell, M., Mocarski, E. and Lehman, I. (1986) Proc. Natl. Acad. Sci. USA **83**, 6322-6326
30. Mackem, S. and Roizman, B. (1980) Proc. Natl. Acad. Sci. USA **79**, 4917-4921
31. Everett, R. (1983) Nuc. Acid. Res. **11**, 6647-6666
32. Peterson, R.C., Doering, J.L. and Brown, D.L. (1980) Cell **20**, 131-141
33. Darby, M.K., Herrera, R.E., Vosberg, H.-P. and Nordheim, A. (1986) EMBO J. **5**, 2257-2265
34. Ackerman, P., Glover, C.V. and Osheroff, N. (1985) Proc. Natl. Acad. Sci. USA **82**, 3164-3164
35. Zenke, M., Grundstrom, T., Matthes, H., Wintzerith, M., Schatz, C., Wildeman, A. and Chambon, P. (1986) EMBO J. **5**, 387-397
36. Maniatis, T., Goodbourn, S. and Fischer, J.A. (1987) Science **236**, 1237-1244
37. Xiao, J.H., Davidson, I., Ferrandon, D., Rosales, R., Vigneron, M., Macchi, M., Ruffenach, F. and Chambon, P. (1987) EMBO J. **6**, 3005-3013
38. Hesse, J.E., J.M. Nickol, M.R. Lieber, and G. Felsenfeld. (1986) Proc. Natl. Acad. Sci. USA **83**:4312-4316
39. Salzman, N.P., Natarajan, V., and Selzer, G.B. (1986) In Salzman, N.P. (ed), The Papovaviridae, Volume 1, The Polyomaviruses, Plenum Press, New York, Vol 1, pp.27-98.
40. Lang, J.C., Spandidos, C. and Wilkie, N.M. (1984) EMBO J. **13**, 389-395
41. Kollias, G., Hurst, J., deBoer, E. and Grosveld, F. (1987) Nuc. Acid. Res. **15**, 5739-5747
42. Sander, M. and Hsieh, T. (1985) Nuc. Acid. Res. **13**, 1057-1072
43. Queen, C. and Baltimore, D. (1983) Cell **33**, 741-748

44. Pommier, Y., Covey, J., Kerrigan, D., Mattes, W., Markovits, J. and Kohn, K.W. (1987) Biochem. Pharmacol. **36**, 3477-3486
45. Hearing, P. and Shenk, T. (1983) Cell 33, 695-703
46. Raisback, G. (1964) Inormation Theory: An Introductioi for Scientists and Engineers. MIT press, Cambridge, MA
47. Harr, R., Haggstrom, M., Gustafsson, P. (1983) Nuc. Acid. Res. **11**, 2943-2957
48. Stadem. R. (1984) Nuc. Acid. Res. **12**, 505-519
49. Mulligan, M., Hawley, D., Entriken, R. and McClure, W. (1984) Nuc. Acid. Res. **12**, 789-801
50. Schneider, T., Stormo, G., Gold, L. and Ehrenfeucht, A. (1986) J. Mol. Biol. **188**, 415-431
51. Stormo, G. (1988) Ann. Rev. Biophys. Biophys. Chem. **17**, 241-263