INFORMATION TO USERS

The most advanced technology has been used to photograph and reproduce this manuscript from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Characterization of recognition elements for DNA cleavage by topoisomerase II

Spitzner, Jeffrey Richard, Ph.D.

The Ohio State University, 1990
CHARACTERIZATION OF RECOGNITION ELEMENTS FOR DNA CLEAVAGE BY TOPOISOMERASE II

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of the Ohio State University

By

Jeffrey Richard Spitzner, B.S.

* * * * *

The Ohio State University

1990

Dissertation Committee: Approved by

Mark T. Muller, Ph.D.
C. William Birky, Ph.D.
Lee F. Johnson, Ph.D.

Adviser
Molecular, Cellular, and Developmental Biology Program
ACKNOWLEDGMENTS

I want to express sincere gratitude to my adviser, Dr. Mark Muller, for his continuing guidance throughout the scientific development of both my project and myself. I also thank Dr. Philip Perlman and the other members of the Molecular, Cellular, and Developmental Biology Program for their assistance during my graduate training. My work would not have been possible without the enduring support of my parents, and I am particularly indebted to my father for supplying his expertise in data analysis. Finally, I wish to thank all who have contributed to my personal development, as science can be detached, but not ultimately removed from the scientist.
VITA


1984 . . . . . . . . . . . . . . . . . . . . . . . . B.S., Ohio State University, Columbus, Ohio

1990 . . . . . . . . . . . . . . . . . . . . . . . . Ph.D., Ohio State University, Columbus, Ohio

PUBLICATIONS


FIELD OF STUDY

Major Field: Molecular, Cellular, and Developmental Biology
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>ACKNOWLEDGEMENTS</th>
<th>ii</th>
</tr>
</thead>
<tbody>
<tr>
<td>VITA</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>ABBREVIATIONS</td>
<td>ix</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>A. Topoisomerases</td>
<td>1</td>
</tr>
<tr>
<td>B. Isolation of type II topoisomerases and genes</td>
<td>3</td>
</tr>
<tr>
<td>C. Enzymatic activity of topoisomerase II</td>
<td>5</td>
</tr>
<tr>
<td>D. Biological roles of topoisomerase II</td>
<td>7</td>
</tr>
<tr>
<td>E. Drugs that inhibit topoisomerase II</td>
<td>11</td>
</tr>
<tr>
<td>F. Topoisomerase II sites in DNA</td>
<td>14</td>
</tr>
</tbody>
</table>

## CHAPTER

### I. PURIFICATION AND CHARACTERIZATION OF TOPOISOMERASE II FROM CHICKEN ERYTHROCYTES

Introduction  
Materials and methods  
Results  
Discussion

### II. GENERATION OF MONOCLONAL ANTIBODIES AGAINST CHICKEN TOPOISOMERASE II

Introduction  
Materials and methods  
Results  
Discussion

iv
LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Activity of topo II fractions through purification</td>
<td>32</td>
</tr>
<tr>
<td>2. DNA substrates used to derive consensus sequence</td>
<td>66</td>
</tr>
<tr>
<td>3. Consensus sequence information must be recovered by alignment</td>
<td>69</td>
</tr>
<tr>
<td>4. Nucleotide frequencies used to derive the topoisomerase II consensus</td>
<td>75</td>
</tr>
<tr>
<td>5. Frequency of topoisomerase II consensus in enhancers</td>
<td>87</td>
</tr>
<tr>
<td>6. Nucleotide proportions in the topoisomerase II consensus sequence</td>
<td>104</td>
</tr>
<tr>
<td>7. Homology calculations at a sample site</td>
<td>106</td>
</tr>
<tr>
<td>8. Comparison of the symmetric and asymmetric consensus sequences to the data base</td>
<td>113</td>
</tr>
<tr>
<td>9. Unique score is the better measure of topoisomerase II cleavage strength than the combined score</td>
<td>116</td>
</tr>
<tr>
<td>10. Comparison of data bases</td>
<td>129</td>
</tr>
<tr>
<td>11. Recognition sites and homology scores in pRYG</td>
<td>138</td>
</tr>
<tr>
<td>12. Comparison of Drosophila topo II sites with the chicken topo II consensus sequence information</td>
<td>151</td>
</tr>
<tr>
<td>13. Base proportions of the aligned Drosophila topo II data set</td>
<td>155</td>
</tr>
<tr>
<td>14. Base proportions of the composite (chicken and fly) pooled data set after alignment</td>
<td>158</td>
</tr>
<tr>
<td>15. Comparison of the composite data set consensus sequence with the chicken topo II consensus sequence</td>
<td>160</td>
</tr>
</tbody>
</table>
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Purification of topo II from chicken erythrocytes.</td>
<td>26</td>
</tr>
<tr>
<td>2. BioRex 70 column elution/activity profile.</td>
<td>28</td>
</tr>
<tr>
<td>3. Phenyl-Sepharose column activity profile and titration of purified topo II.</td>
<td>30</td>
</tr>
<tr>
<td>4. Titration of topo II relaxation +/- ATP.</td>
<td>31</td>
</tr>
<tr>
<td>5. Topo II is a homogeneous band by SDS-polyacrylamide gel electrophoresis and by western blot analysis.</td>
<td>34</td>
</tr>
<tr>
<td>6. Inhibition of topo II activity by several reagents</td>
<td>36</td>
</tr>
<tr>
<td>7. Titration of high activity topo II preparation.</td>
<td>38</td>
</tr>
<tr>
<td>8. Preliminary western blot analysis of antibodies using a whole cell MSB-1 extract.</td>
<td>49</td>
</tr>
<tr>
<td>9. Western blot analysis of hybridomas against a whole cell extract and against purified topo II.</td>
<td>51</td>
</tr>
<tr>
<td>10. Western blot analysis of anti-topo II antibodies using radioiodinated Protein A.</td>
<td>53</td>
</tr>
<tr>
<td>11. Analysis of nucleotide frequency.</td>
<td>71</td>
</tr>
<tr>
<td>12. Analysis of redundancy in the data base.</td>
<td>73</td>
</tr>
<tr>
<td>13. Topo II cleavage on HSV-1 promoter fragment.</td>
<td>79</td>
</tr>
<tr>
<td>14. Analysis of nucleotide frequency and redundancy that defines a symmetric consensus sequence.</td>
<td>82</td>
</tr>
<tr>
<td>15. Comparison of symmetric and asymmetric topo II consensus sequences to actual sites.</td>
<td>115</td>
</tr>
<tr>
<td>16. Topoisomerase II cleavage on an HSV-1 fragment is predicted by the unique score model.</td>
<td>118</td>
</tr>
</tbody>
</table>

vii
17. The unique score model makes accurate predictions of topoisomerase II cleavage sites and frequencies in the absence or presence of m-AMSA.
18. Topo II cleavages on an archaebacterial sequence.
19. Optimizing discrimination between sites, nonsites.
20. Topoisomerase II cleavage strength correlates significantly with the unique score.
22. Comparison of topoisomerase II cleavages from different organisms.
23. Redundancy of composite data set, by position relative to the cleavage site.
25. Time dependence for release of decatenated kDNA from low melt agarose inserts.
26. Topo II decatenates kDNA from total Crithidia fasciculata DNA purified in agarose inserts.
27. Quantitative use of the gel release assay.
ABBREVIATIONS

AS .................. Ammonium sulfate
ATP .................. Adenosine triphosphate
BSA .................. Bovine serum albumin
C.fasciculata ...... Crithidia fasciculata
Dal .................. Dalton
DNA .................. Deoxyribonucleic acid
Drosophila .......... Drosophila melanogaster
E.coli .............. Escherichia coli
EDTA ................. Ethylenediaminetetraacetic acid
g .................. gram
hr .................. Hour
K .................. Guanine or Thymine bases
kb .................. kilobases
kD, kDal ............. kiloDalton
kDNA ................ Kinetoplast DNA (C.fasciculata)
L .................. Liter
m-AMSA ............... 4'-(9-acridinylamino)methansulfon- m-anisidide
MAR .................. Matrix associated region
MCN .................. Micrococcal nuclease
min .................. Minute
N .................. Any nucleotide (unspecified)
non-"N" ............... nonrandom base frequency
oligo ................. Oligonucleotide
PMSF ................. Phenylmethylsulfonylfluoride
R .................. Purine
SAR .................. Scaffold attachment region
SDS .................. Sodium dodecylsulfate
sec .................. Second
SV40 .................. Simian Virus 40
topo I ................ Topoisomerase I
topo II ................. Topoisomerase II
VM26 .................. teniposide
VP16 .................. etoposide
xg (centrifugation) .. times gravities of acceleration
X.laevis ............. Xenopus laevis
Y .................. Pyrimidine
INTRODUCTION

A. Topoisomerases

The nature of DNA is such that its topological properties impinge greatly upon its cellular functioning. These topological properties of DNA are regulated by enzymes known as topoisomerases; the reaction characteristic to all topoisomerases (topos) is the ability to break and then reseal the DNA sugar-phosphate backbone. Topos are essential for providing the conformational adjustments of DNA that are associated with basic cellular phenomena such as transcription, replication, and chromosome organization. As a result, all organisms (yet examined) and some viruses encode genes for topos.

The first topoisomerase identified was the *E. coli* omega-protein that removed negative superhelical turns from supercoiled DNA (Wang, 1971). A eukaryotic counterpart to this protein was soon found in embryonic mouse cells (Champoux and Dulbecco, 1972) and initially called the untwisting enzyme; this enzyme catalyzed the relaxation of either negatively or positively supercoiled DNA. Now, both of these proteins have been combined into a class of enzymes
known as type I topoisomerases. The general properties of topo I enzymes are that they make transient single-strand breaks of DNA and change the linking number of DNA in steps of one in a reaction that does not require any energy cofactors such as ATP.

A second class of enzymes is known as type II topoisomerases; these enzymes make transient double strand breaks in DNA through which an intact duplex passes to alter the linking number in steps of two in a reaction requiring ATP. The first topo II identified was the *E.coli* protein DNA gyrase that introduces negative supercoils into relaxed closed circular DNA molecules (Gellert, 1976 (a)). Type II topoisomerases were also identified in eukaryotes (Baldi et al, 1980; Miller et al, 1981). In contrast to the prokaryotic enzyme, these topo II's catalyzed only the relaxation of supercoiled DNA. Several other topos have been identified which do not fit clearly into either of the general two classes; among these is a reverse gyrase that introduces positive supercoils into DNA (Kikuchi and Asai, 1984), topoisomerase II' from *E.coli*, a topo II that does not require ATP (Brown et al, 1979; Gellert et al, 1979), and *E.coli* topo III, which has a type I topo activity but cation requirements different from that of the omega-protein (Dean et al, 1983; Srivenugopal et al, 1984). Quite a bit of knowledge has been collected about the structure and function of topoisomerases, and extensive details have been
presented in reviews by Vosberg (1985) and by Wang (1985); therefore, the focus within will be restricted to eukaryotic type II topoisomerases.

B. Isolation of type II topoisomerases and their genes.

An ATP-dependent DNA catenation activity was first demonstrated with a cell extract from X. laevis oocytes (Gandini-Attardi et al, 1976) and also shown by Liu et al (1980). Topo II was purified from Xenopus laevis germinal vesicles (Baldi et al, 1980), Drosophila melanogaster embryos (Hsieh and Brutlag, 1980; Sander and Hsieh, 1983; Shelton et al, 1983), HeLa cells (Miller et al, 1981), Yeast (Goto and Wang, 1982; Goto et al, 1984), rat liver (Duguet et al, 1983), and calf thymus (Darby and Vosberg, 1985; Halligan et al, 1985), and has since been purified from a number of additional sources. The pattern that has emerged from all of these type II topoisomerases studied is that their properties appear to be highly conserved between different organisms. In all cases, the enzyme appears to function as a homodimer, with an average monomer molecular weight of 170,000 kD (although smaller polypeptides, most likely proteolytic fragments, are also observed). In addition, topo II is generally found in the cell in
relatively large copy numbers (100,000 to 1,000,000 molecules per cell in rapidly growing cells, Miller et al, 1981; Earnshaw and Heck, 1988; Hsiang et al, 1988). Topo II's from different species are functionally very similar both in vitro and in vivo (see above and below).

The nucleotide sequence has been determined for several topoisomerase II genes, and these genes have been mapped to specific chromosomes. The complete coding sequences have been published for the topo II genes of Saccharomyces cerevisiae (Giaever et al, 1986), Schizosaccharomyces pombe (Uemura et al, 1986), D.melanogaster (Wyckoff et al, 1989) and humans (Tsai-Pflugfelder et al, 1988) and these genes demonstrate about 50% amino acid sequence homology between one another. There is also some homology between these eukaryotic topo II enzymes and bacteriophage T4 topo II (Huang, 1986a,b) and bacterial DNA gyrase (Moriya et al, 1985; Adachi et al, 1987; Swanberg and Wang, 1987). Genetic analyses have suggested that eukaryotic topo II molecules are functionally conserved; the S.pombe topo II gene can complement a mutant S. cerevisiae gene (Uemura et al, 1986), as can the D.melanogaster topo II gene (Wyckoff and Hsieh, 1988). While it appears that the yeasts code for only a single topo II gene, a recent report by Chung et al (1989) has suggested that human cells may have two different, but highly homologous topo II genes.
C. Enzymatic activity of topoisomerase II.

As stated above, these enzymes catalyze ATP-dependent relaxation of negative and positive supercoils in DNA; type II topos can also knot, unknot, catenate, and decatenate DNA molecules (see Vosberg, 1985 and Wang, 1985 for reviews). The interaction between topo II and a binding site seems to consist of two basic steps (see Vosberg, 1985; Wang, 1985; Liu, 1989, and references within for the details of what follows). First, the enzyme binds electrostatically to a site in a manner such that treatment with a protein denaturant, such as sodium dodecylsulfate (SDS), leads to dissociation of the complex. The second step involves breakage of phosphodiester bonds in the sugar-phosphate backbone of the DNA and a covalent linkage between the protein and DNA (known as a "cleavable complex"); treatment of this complex with SDS results in DNA breaks with topo II monomers linked via tyrosine residues to 5' phosphates of the DNA. The active tyrosine of yeast topo II has been identified (Worland and Wang, 1989). Both single and double-strand breaks are observed (at generally identical sites, Muller et al, 1988; Lee et al, 1989), and it appears that the topo II dimers act concertedly, but sequentially, so that SDS might trap a transitional event in the formation of the double-strand gate through which a DNA duplex would pass. As with bacterial DNA gyrase, the
double-strand break leaves a four base 5' overhang; the 3' ends have free hydroxyl groups.

A variety of assays have been used to measure the enzymatic activity of topo II in a sample. Certain assays make use of topological conversions of DNA that can be analyzed by agarose gel electrophoresis. These include relaxation of supercoiled plasmid DNA (Keller and Wendel, 1974), unknotting of knotted phage P4 DNA (Liu et al, 1981), and decatenation of the interlocked networks of DNA that are isolated from the mitochondria of certain trypanosomes (Marini et al, 1980); the presence of topo I interferes with a relaxation assay because it also catalyzes this reaction (Gellert et al, 1976); however, the other reactions are unique to topo II. Alternative methods for measuring topo II activity include precipitating covalent protein-radiolabeled DNA complexes, followed by scintillation counting of samples (Trask et al, 1984), and monitoring the cleavage of a radiolabeled DNA substrate by gel electrophoresis (Sander and Hsieh, 1983; Liu et al, 1983).

Extensive characterization of the biochemical reactions catalyzed by topo II have been reported (in essentially all the papers cited above, and reviewed by Vosberg, 1985). In brief, magnesium ions are required for formation of a covalent complex with DNA (Sander and Hsieh, 1983; Trask et al, 1984; Udvardy et al, 1986) and ATP further stimulates complex formation (Udvary et al, 1986). Osheroff et al
(1983) found that under optimal reaction conditions the relaxation activity is processive, and that binding of ATP is required for strand passage while hydrolysis of ATP is required for enzyme turnover. The effects of other ions on topo II have also been investigated (see, for example, Osheroff, 1987; Osheroff and Zechiedrich, 1987; Zwelling et al, 1988; Lee et al, 1989). There is also evidence that topo II activity is modified by other cellular factors (Goto and Wang, 1982; Glikin et al, 1984; Darkin and Ralph, 1989) and by posttranslational modifications to the enzyme itself (for example, by phosphorylation, see below).

D. Biological roles of topoisomerase II

In vitro, topo II can relieve the torsional stress on DNA and unlink intertwined DNA molecules; it was assumed that topo II would supply these activities within cells, as well. Replication of the yeast 2 micron plasmid (DiNardo et al, 1984) or of SV40 (Yang et al, 1987) in the absence of topo II activity resulted in catenated dimers. Sundin and Varshavsky (1981) were the first to suggest that topo II might be required at the end of DNA replication to allow passage of one duplex through the other and subsequent separation of daughter chromosome molecules. Analysis of
temperature-sensitive mutations in the topo II genes of *Saccharomyces cerevisiae* (DiNardo et al, 1984; Holm et al, 1985) and *Schizosaccharomyces pombe* (Uemura and Yanagida, 1984) showed that topo II is an essential cellular protein that is required for the disjunction of sister chromatids at anaphase; at non-permissive temperatures, cells underwent one round of DNA replication, then were blocked at the stage of nuclear division and soon died. Further genetic analysis suggested that topo II can only disjoin sister chromatids at the onset of anaphase, as cells died even if shifted back to the permissive temperature (Holm et al, 1989). The results of the analysis of mitosis in *S.pombe* conditional topo II mutants (Uemura et al, 1987) and of using topo II inhibitors in an in vitro mitotic induction system (Newport, 1987) both implied that topo II is also required for cells to attain complete chromosome condensation. Further studies with topo mutants of yeasts have suggested that topo II can also function within cells to relieve the torsional stress on DNA that is generated by transcription (Uemura and Yanagida, 1984; Brill et al, 1987). Additionally, topo II can act as a swivel for DNA replication (Yang et al, 1984; Snapka et al, 1988), and was found to associate directly with newly replicated DNA (Nelson et al, 1986; Woynarowski et al, 1988). Finally, topo II has been implicated in recombination (Bae et al, 1988; Sperry et al, 1989) and in DNA supercoiling (Glikin et al, 1984).
In addition to its enzymatic functions, topo II also has a structural role within cells. Use of antibodies against topo II showed that it was a major component of the mitotic chromosome scaffold (Earnshaw et al, 1985; Gasser et al, 1988) and is preferentially located at the base of chromatin loop domains (Earnshaw and Heck, 1985). Topo II is also a major component of the interphase nuclear matrix (Berrios et al, 1985). It appears that there are specific regions of DNA that associate with the nuclear scaffold (Mirkovitch et al, 1984), and that these are identical to the sites of interaction with the nuclear matrix (Cockerill and Garrard, 1986). There is evidence that the regions of DNA that associate with the nuclear scaffold and matrix contain topo II binding sites (Cockerill and Garrard, 1986; Gasser and Laemmli, 1986; Sperry et al, 1989). It is tempting to speculate that, since topo II has been localized to the nuclear scaffold/matrix, and DNA makes contact with the nuclear scaffold/matrix at specific attachment regions, topo II would interact with these attachment regions in cells; however, there is no experimental proof for this theory. The idea that topo II itself was responsible for chromosome loop organization by fastening attachment sites to the scaffold/matrix has been abandoned because attachment sites appear to be cell cycle independent (Cockerill and Garrard, 1986), while topo II levels are highly cell cycle regulated (Heck et al, 1988; Heck et al, 1989).
A number of reports suggested that topo II levels were greater in proliferating cells than in quiescent cells (Duguet et al, 1983; Zwelling et al, 1987; Nelson et al, 1987; Heck and Earnshaw, 1986). Quantitative immunological methods were used by Heck et al (1988 and 1989) to show that topo II levels parallel the chromosome condensation/decondensation cycle; levels were found to begin increasing at the end of the onset of S phase and peak during mitosis, followed by a drastic decrease in enzyme stability and a several-fold loss of total topo II. It was found that topo II activity increased at a much greater rate during the cell cycle than did the actual enzyme concentration (Estey et al, 1987), suggesting that topo II was being modified. It has been shown that topo II can be phosphorylated in vitro, leading to an increase in enzymatic activity (Ackerman et al, 1985; Sahyoun et al, 1986) and that in the sponge Geodia, agents that stimulate cell proliferation caused in vivo phosphorylation of topo II (Rottman et al, 1987). Phosphorylation of topo II in vivo was observed in Drosophila (Ackerman et al, 1988) and mouse (Saijo et al, 1990) cells, and Heck et al, (1989) demonstrated that out of a number of antigenically related polypeptides found in chicken lymphoblast cells, the highest molecular weight (170 kD) topo II species was specifically phosphorylated in vivo during mitosis. They suggested that this form is involved in disjunction of sister chromatids at
anaphase, while the other forms (some of which were present in large quantities) may have different cellular functions. It has also been reported that endogenous topo II levels are developmentally regulated within cells (Fairman and Brutlag, 1988; Roca and Mezquita, 1989; Luke and Bogenhagen, 1989).

E. Drugs that inhibit topoisomerase II.

Many antitumor drugs interact with DNA (Waring, 1981); Filipski (1983) suggested that topoisomerases might be the cellular targets for some of these drugs, particularly DNA intercalators. Many studies have shown that prokaryotic and eukaryotic topoisomerases are the targets through which a number of drugs mediate their cytotoxicity, as reviewed by Wang (1985), Vosberg (1985), Drlica and Franco (1988), Rose (1988), Pommier and Kohn (1989) and Liu (1989). Since a large body of work is already available covering these drugs and their potential value for use in anticancer chemotherapy, only drugs of particular relevance to the experiments to be presented within will be discussed here.

The addition of the potent antitumor drug 4'-{(9-acridinylamino)methansulfonyl- methyl}anisidide (m-AMSA) to cells was found to induce DNA breaks with protein tightly linked
to the broken ends (Ross and Bradley, 1981; Zwelling et al, 1981). Ross et al (1979) had suggested that a topoisomerase might play a role in these breaks, and further studies with m-AMSA or with another agent, ellipticine, showed that these drugs led to a stimulation of topo II-DNA complexes in experiments using purified topo II (Nelson et al, 1983; Tewey et al, 1984). Both single and double-strand breaks were identified in the DNA, with topo II bound to the 5' ends of the breaks via a phosphotyrosyl linkage (Rowe et al, 1986). An isomer of m-AMSA, o-AMSA, that does not have antitumor activity and does not induce DNA breaks also does not interfere with topo II activity in vitro (Nelson et al, 1984). m-AMSA induced damage is reversible; if prior to terminating reactions with SDS, high salt is added, or if the drug is diluted out, then the DNA cleavage is reduced (Pommier et al, 1983; Tewey et al, 1984). Nelson et al (1984) suggested that m-AMSA acted by trapping the covalent topo II-DNA intermediate by forming a ternary drug-enzyme-DNA complex that was much more stable than the normal enzyme-DNA complex that reseals rapidly.

Evidence from a number of experiments has suggested that m-AMSA mediates its cellular effects strictly through interaction with topo II. SV40 chromatin studied after drug treatment of infected cells revealed formation of nicked and linear viral DNA molecules, and all of the DNA breaks were covalently bound to protein (Yang et al, 1985); furthermore,
these damaged DNA molecules were selectively immunoprecipitated by use of an antibody specific for topo II. In addition, the drug stimulation of purified topo II cleavage of DNA was shown to be reversible and not sensitive to DNA conformation (and therefore not caused just by intercalation of the drug, Nelson et al, 1984; Tewey et al, 1984a; Tewey et al, 1984b). Consistent with the model for m-AMSA induced DNA damage, X-ray irradiation of cells did not lead to more protein linked DNA breaks in drug treated cells (Pommier et al, 1984), and there was found to be a very good correlation between the number of protein linked DNA breaks and the cytotoxicity of the drug (Rowe et al, 1986; Covey et al, 1988), suggesting that m-AMSA killed cells strictly through stimulation of topo II cleavage of DNA. The drug induced DNA damage of cells is proportional to endogenous levels of topo II activity (which is why proliferating cells, particularly tumor cells, are selectively killed; see above reviews), and m-AMSA induced DNA breaks also cause chromosome aberrations and sister chromatid exchanges (Andersson and Kihlman, 1989).

A number of other drugs are thought also to cause breakage of cellular DNA by trapping topo II-DNA covalent complexes, including other DNA intercalators, such as ellipticine derivatives, doxorubicin (Adriamycin), daunamycin (see reviews by Rose, 1988; Drlica and Franco, 1989; Pommier and Kohn, 1989; Liu, 1989), and Amonofide
Additionally, there are nonintercalative antitumor drugs that also mediate their cytotoxicity through stimulation of topo II linked cleavage of DNA. The epipodophyllotoxins VM26 (teniposide) and VP16 (etoposide) are the best studied topo II inhibitors of this type (see reviews above for details). These drugs also inhibit the strand passing activity of topo II and have essentially the same effects upon cells and topo II as m-AMSA. In addition to their clinical value for anticancer chemotherapy, topo II inhibitors have also proved valuable for stimulating topo II cleavage of DNA to aid in mapping topo II sites (see below).

F. Topoisomerase II sites in DNA.

As stated previously, DNA breakage is an intermediate in the reaction mechanism of topo II. Interruption of the reaction by a protein denaturant such as SDS traps a small percentage of the reactants as covalent topo II-DNA complexes, with protein linked to the 5' ends of the DNA at the break sites. These DNA cleavages are a permanent record of where topo II was acting, and these topo II binding sites can be mapped. The evidence suggests that the cleavage sites are indeed the sites of catalytic activity (Sander et
al, 1987). Mapping topo II sites entails termination of reactions with protein denaturants, purification of the DNA (including protease treatment to digest the covalently attached protein), and restriction enzyme digestion if desired, followed by electrophoresis to separate the uncleaved parental DNA and the cleaved subfragments that will have greater mobility on the gels. If native (neutral buffer) gels are run, then only the double-strand breaks will be observed; if the DNA is analyzed under denaturing conditions, both single and double-strand breaks will be identified. The breaks can be mapped on large scale (cleaved bands 300 to 20,000 bases in length with sites identified to within about 5% error) by agarose gel electrophoresis, using a gel concentration appropriate for the DNA size expected, or sequencing gels can be used to identify topo II sites at single base resolution (however, sites have been sequenced only on small fragments, less than 500 bases). If the DNA is radioactive (usually $^{32}$P-end labeled for mapping sites), the cleaved subbands can be observed after exposing the gel directly against X-ray film. If the DNA is not radiolabeled then the DNA must be transferred to a membrane (nitrocellulose or nylon) and hybridized with a specific radioactive probe (for details see the references below). Thus from the mobilities of the topo II induced subbands (compared to the appropriate size markers or sequencing ladders), the sites of topo II
activity can be deduced.

Initial characterizations of topo II sites (using purified topo II) showed that the enzyme cleaved DNA at many sites, and the sites were reproducible (not random), but there was no obvious sequence homology observed (Sander and Hsieh, 1983; Liu et al, 1983); similar results were reported for the analysis of DNA sites cleaved by bacterial DNA gyrase (Morrison and Cozzarelli, 1979). Udvardy et al (1985) mapped Drosophila topo II sites in vitro on a plasmid containing a 10 Kb insert of a Drosophila heat shock locus and flanking sequences. Surprisingly, they found a very non-random distribution of sites; intense cleavages were observed in the intragenic regions (between two heat shock genes) while there was little cleavage in coding regions. The strong cleavage bands were closely correlated with the locations of nuclease hypersensitive sites. They also reported strong topo II cleavage sites in only the intragenic regions of a plasmid containing a Drosophila histone gene cluster; these sites matched the micrococcal nuclease hypersensitive sites on the plasmid. In contrast, Micrococcus luteus DNA gyrase showed a very different cleavage pattern. Udvardy et al (1985) suggested that these sites might have significance in vivo as well, as the nuclear matrix attachment sites for both of the gene clusters had been mapped (Mirkovitch et al, 1984) and correspond to the locations of the topo II site clusters.
Darby et al (1986) examined calf thymus topo II cleavage of human and mouse c-fos genes with overlapping DNA fragments (that could be analyzed on sequencing gels) and found that cleavage sites were located in important regulatory regions (that are also nuclease hypersensitive sites) rather than in coding sequences. They showed that inclusion of the antitumor drugs VP16, VM26, or m-AMSA stimulated the topo II cleavages at essentially the same sites as in their absence. ATP also enhanced the cleavage efficiency. In addition, a HeLa cell nuclear extract that contained topo II cleaved the DNA at the same sites as the calf thymus enzyme.

Sander and Hsieh (1985) sought to study the sequence specificity in more detail, and mapped sites of *Drosophila* topo II on a plasmid containing a *Drosophila* P element and flanking sequences. Regions in which strong double strand topo II cleavage sites were found were then analyzed on sequencing gels. They identified 16 strong topo II cleavage sites on the A+T rich sequence and derived a consensus sequence from these sites. Eleven conserved base positions were identified, all closely flanking the cleavage site. The consensus sequence is degenerate and no positions were absolutely conserved within the sites. The strong sites showed an average 60% match to their consensus sequence, while weak sites identified shared only 50% homology with the consensus sequence. The consensus sequence was not very
useful for predicting cleavage sites; of sites with greater than 55% consensus match, only one in four was actually cleaved by topo II. However, it was shown by Lee et al (1989) that a small cloned oligonucleotide containing a 100% match to the Drosophila topo II consensus sequence was still a strong cleavage site. Additionally, they performed DNaseI footprinting experiments and showed that topo II protects a region of about 25 bases, centered around the cleavage site. This further suggests that cleavage sites are also the binding sites for the enzyme, and indicates that the sequence-specific recognition elements for topo II are found around the cleavage site.

The mapping of topo II sites in vivo has required the use of drugs both because the trapping of covalent topo II-DNA complexes is too low for efficient detection of cleaved molecules, and because the stimulation of cleavage at a site by a topo II inhibitor demonstrates that the break was caused by topo II and not some other enzyme. Yang et al (1985) mapped topo II sites in SV40 chromatin in the presence of m-AMSA. They identified a single major cleavage site that was preferentially induced at late times in infection and mapped to the nuclease hypersensitive control region of the virus. The major site in vivo was observed as a minor band in vitro, suggesting that access to the DNA or structural conformation of chromatin is more important in regulating topo II sites in vivo than simply DNA sequence.
Udvardy et al (1986) mapped topo II cleavages in Drosophila nuclei in the presence of the topo II inhibitor VM26. They first showed that the cleavages in vitro in the presence of VM26 were roughly the same as without the drug, namely in the intragenic spacer 5' of the gene, and at the 5' and 3' ends and flanks of the gene. In nuclei, the endogenous topo II cleaved DNA intensely at the 5' end of the gene (an untranslated region) and in the intragenic spacer; The strongest sites corresponded to the locations of major micrococcal nuclease hypersensitive sites. Activation of the gene by heatshocking the cells before isolating nuclei led to a somewhat altered cleavage pattern (and it is known that heat shock induces changes in chromatin structure). When Udvardy et al (1986) mapped topo II sites in the histone gene cluster in vivo (in nuclei, as above), they found cleavage sites only in the intragenic regions, paralleling strongly both the sites of topo II cleavage in vitro and the sites for micrococcal nuclease hypersensitivity in vivo.

Rowe et al (1986) also mapped the VM26 induced topo II cleavage sites in the Drosophila heat shock gene, but in intact cells. Their results were quite similar to those of Udvardy et al (1986); topo II cleaved strongly in the intragenic spacer 5' of the gene and at the 5'end of the gene. These sites corresponded well to the DNase I hypersensitive sites in nuclei. Again, heat shock altered
the cleavage pattern somewhat, with more cleavage within the coding region. A study by Muller and Mehta (1988) demonstrated in vivo cleavages on the βA-globin gene of embryonic chicken erythrocytes in the 5' nucleosome-free region, in the second intron, and in the 3' enhancer element. The topo II sites in vivo were a subset of those observed in vitro, and the sites matched a number of the in vivo DNaseI hypersensitive sites; however, further studies showed that topo II cleavage and nuclease hypersensitivity were, in fact, separable phenomena. Collectively, the results from mapping topo II sites in chromatin suggest that topo II acts primarily at noncoding regions (enhancers and promoters), probably at exposed (i.e. accessible, nucleosome-free) sites on the DNA helix. It is not certain whether the sites identified are merely incidental (real, but without cellular importance), exist only for mitotic functions, or have other cellular functions. Only further experimentation will allow elucidation of topoisomerase II activity within cells.
Chapter I

Purification and characterization of topoisomerase II from chicken erythrocytes.

INTRODUCTION

The best way to study a particular enzyme is to begin by obtaining it in a purified form. Rather than rediscovering an enzyme that had already been purified, it was felt that more could potentially be learned from a different topo II. Chicken topo II had been shown to associate with metaphase chromosome scaffolds in cells (Earnshaw and Heck, 1985), but it had never been purified. Rather than using chicken tissues (which are difficult to work with) or culturing a chicken cell line (which is expensive), chicken erythrocytes were selected as the source. It is inexpensive to collect large quantities (liters) of chicken red blood cells, and these cells are low in proteolytic activity. Topo I was previously purified from this source (Trask and Muller, 1983), and nuclear extracts were shown by decatenation of kDNA to contain topo II as well. Chicken erythrocytes are relatively poor in topo II levels (Earnshaw and Heck, 1985), but cell
concentrations are a thousand-fold higher than cultured
cells attain; in addition, erythrocytes appear to express
fewer proteins than MSB-1 cells (a chicken lymphoblast cell
line; unpublished results), so that there are fewer
contaminating activities to purify away from topo II.

Type II topoisomerases have been purified from a number
of sources (see references above); however, all published
procedures were lengthy and arduous, requiring 4-7 column
chromatography steps, and dialysis or glycerol gradient
steps. All of these contribute to loss of activity because
topo II is not a very stable enzyme. Therefore, a novel
rapid and inexpensive purification scheme was developed and
topo II was purified from chicken erythrocytes and
characterized. It was found to be homologous in structure
and activity to the other eukaryotic type II topoisomerases,
as anticipated.
MATERIALS and METHODS

Materials. Chemical reagents and Phenyl-Sepharose were from Sigma, BioRex 70, acrylamide, and protein assay kit were from BioRad, and oligonucleotides were from Pharmacia. Nitrocellulose was from Schleicher and Schuell.

Chromatography matrix preparation. BioRex 70 was washed twice in 1 M KOH prior to equilibration in BR buffer (0.2 M potassium phosphate pH 7.1, 10% glycerol, 25 mM 2-mercaptoethanol, 0.5 mM PMSF). Phenyl-Sepharose was degassed then equilibrated in PS buffer (20% ethylene glycol, 25 mM potassium phosphate pH 7.1, 25 mM 2-mercaptoethanol, 0.5 mM PMSF).

DNA substrates. Plasmid pXP1 (pSP64 containing the Xenopus 5sRNA gene, provided by D. Brown) was isolated by the alkaline lysis method (Maniatis et al, 1982). Kinetoplast DNA (kDNA) was purified from C.fasciculata (provided by P. Englund) as described by Saucier et al (1981) with the modifications of Hajduk et al (1984) and resuspended in TE (10 mM Tris-HCl pH 7.5, 1 mM EDTA) at a concentration of 250 ng/μL.
**Enzyme assays.** The standard decatenation assay was a 20 µL reaction in topo II reaction buffer (50 mM Tris-HCl pH 8.0, 120 mM KCl, 10 mM MgCl₂, 0.5 mM ATP, 0.5 mM dithiothreitol, 30 µg/µL bovine serum albumin) with 250 ng kDNA and enzyme added as specified. Reactions were incubated 5-30 min at 30°C then terminated with 5 µL stop dye (5% sarkosyl, 0.025% bromphenol blue, 25% glycerol) and loaded onto a 1% agarose gel and run in TAE gel buffer (Maniatis et al, 1982) containing 0.5 µg/mL ethidium bromide at 7-9 volts/cm. Relaxation reactions were conducted as described in Trask and Muller (1983). SDS/K+ analysis of covalent protein-DNA complexes was performed as described in Trask et al, (1984).

**Protein analysis.** Protein concentrations were determined using the BioRad protein micro assay as recommended. SDS-polyacrylamide gels were run using the discontinuous buffer system described by Laemmli, and proteins were stained with Coomassie Blue (0.5% in 25% trichloroacetic acid). Proteins were transferred to nitrocellulose in a Hoeffer transfer unit in 192 mM glycine, 25 mM Tris base for 3 hr at 0.8 milliamperes, then the membrane was blocked in TBS (20 mM Tris-HCl pH 7.5, 145 mM NaCl) for 2 hr and incubated with antibody overnight at room temperature with shaking. After washing with TBS, blots were incubated with 10 µCi ^125^I- Protein A (ICN) for 2 hr, then washed for several hr in TBS and autoradiographed using Kodak XAR X-ray film.
RESULTS

Purification of topoisomerase II

All steps were conducted at 4°C. Fresh chicken blood was collected at a local slaughter house (Case Poultry, Winesburg, Ohio), transported on ice, and mixed with heparin (20,000 units per liter of blood) to prevent coagulation. The blood was centrifuged for 10 min at 13,000g and the buffy layer (leukocytes) and other debris were removed. The red blood cells were concentrated two-fold and washed twice in TEMP (10 mM Tris-HCl pH 7.6, 1 mM EDTA, 5 mM MgCl₂, 0.5 mM PMSF), then resuspended in TEMP and frozen overnight at -70°C to break the cell membranes (this material can be stored for up to a year without loss of topo II). Two liters of concentrated washed blood were thawed then washed 4 times in TEMP (with 10 min 13,000g centrifugations) to remove hemoglobin and other cytosolic and soluble proteins. The pelleted nuclei were suspended in 800 mL suspension buffer (50 mM Tris-HCl pH 7.6, 50 mM sodium bisulfite (pH 7.6 with Tris base), 30 mM 2-mercaptoethanol, 1
Figure 1. Purification of topo II from chicken erythrocytes. The details are described in the text.
mM PMSF) and solid NaCl was added to a final concentration of 0.3 M (sufficient to extract topo II from nuclei, while topo I, histones, and many other contaminating proteins remain behind). The suspension was stirred for 1 hr, then Polymin P (polyethyleneimine-HCl, pH 7.6) was added to 0.9% and the suspension was stirred for another 30 min. Centrifugation (10 min at 13,000g) removed nucleic acids and pelleted the nuclei, while topo II remained in the supernatant. To this supernatant (650 mL), 118.3 g ammonium sulfate (AS) was added slowly; after stirring 60 min the mixture was centrifuged 20 min at 13,000g. This 33% AS step precipitated many low solubility proteins. Then a 45% AS step (an additional 61.6 g were added) was performed as above; this step precipitated topo II while many higher solubility proteins, such as topo I, remained largely in the supernatant.

The 45% AS pellet was suspended in 10-15 volumes BR buffer with 0.1 M potassium phosphate (KP) (usually 100 mL total) and loaded onto a 2.5 x 30 cm BioRex 70 column preequilibrated in BR buffer with 0.2 M KP. After loading, the column was washed with equilibration buffer until most of the hemoglobin was removed (about 3 column volumes). The column was developed with a 500 mL gradient of 0.2 to 0.6 M KP in BR buffer at 1.5 mL/min and 8 mL fractions were collected. Every second fraction (1 μL each) was assayed for decatenation activity, as shown in Figure 2. Topo II
Figure 2. BioRex 70 column elution/activity profile. The reactions in lanes 1-39 on the gel are standard kDNA decatenation reactions with 250 ng kDNA in topo II reaction buffer. Total volumes are 20 μL, and 1 μL of test column fraction was added. After 15 min incubation at 30°C, 5 μL of stop dye (5% sarkosyl, 0.025% bromphenol blue, 25% glycerol) was added and samples were run on a 1% agarose gel in TAE buffer with 0.5% ethidium bromide at 9 volts/cm for 30 min. Lanes 1-3 are markers: catenated kDNA (kDNA), linear kDNA (L), and control topo II decatenation products (open circular, OC, and relaxed monomer, Rel), respectively (and are indicated in the figure). Lane 4 is the ammonium sulfate fractionation column input, and lanes 5 and 6 are column wash fractions. Lanes 7-39 are assays of every second BioRex 70 column fraction from the 0.2 to 0.6 M potassium phosphate gradient used in the purification (1 μL of the 8 mL fraction). The fractions eluted from this column that had observable topo II activity were lanes 19-33. The fractions from lanes 19-29 had complete decatenation activity, and were pooled together as the topo II BioRex 70 pool.
activity eluted from about 0.35-0.45 M KP$_i$ and these active fractions (lanes 19-29) were pooled as the topo II BioRex 70 pool. This pool was loaded directly onto a 4 mL Phenyl-Sepharose column preequilibrated in PS buffer at 1.5 mL/min. The column was washed with 10 column volumes PS buffer (20% ethylene glycol), then with a 5 column volume step of PS buffer containing 60% ethylene glycol, during which 0.5 mL fractions were collected. Fractions were assayed by decatenation, as shown in Figure 3A. The fractions which completely decatenated the kDNA in 5 min were pooled as the topo II PS pool; this pool was aliquoted and tubes were stored at -70°C, where the enzyme maintains activity indefinitely.

The activity of the PS pool was titrated by serial dilution and assayed by decatenation, as shown in Figure 3B. One unit of topo II activity was defined as the amount of enzyme required to decatenate 50% of 250 ng kDNA in 15 min at 30°C, so the activity of the PS pool was about 10-12 units/µL. The PS pool was shown to have no observable topo I activity, because there was no relaxation activity in the absence of ATP (see Figure 4, lanes 9-14); in the presence of ATP, relaxation was observed, at dilutions up to 1:16.

Table 1 shows the protein concentrations and specific activities of various fractions during the purification. The 33-45% ammonium sulfate step precipitated only a small fraction of the total extracted proteins, and the BioRex
Figure 3. Phenyl-Sepharose column activity profile and titration of purified topo II. The reactions were standard decatenation assays, as in Figure 2. Panel A is the profile of fractions eluted from the Phenyl-Sepharose column. Lane 1 is a linear kDNA marker, lane 2 is catenated kDNA, lane 3 is the column void material, and lanes 4-6 are 20% ethylene glycol wash fractions. Lanes 7-27 are 0.5 mL fractions during the 60% ethylene glycol step. Lanes 15-26 showed considerable topo II activity, so these were combined as the Phenyl-Sepharose topo II pool, to be called purified topo II. The positions of catenated and decatenated kDNA are indicated. Panel B is titration of decatenation activity by the purified topo II. Lanes 1, 2, and 10 are markers, and the decatenated kDNA species are indicated as OC and Rel in the figure. The samples in lanes were serial dilutions of topo II in elution buffer (containing 60% ethylene glycol) with 40 μg/mL bovine serum albumin added. Lanes 3-9 are, respectively, 1:1, 1:2, 1:4, 1:8, 1:16, 1:32, and 1:64 dilutions of 1 μL of topo II.
Figure 4. Titration of topo II relaxation +/- ATP. Purified topo II was diluted as in Figure 3 and added to 200 ng supercoiled pXPl plasmid DNA in 20 µL reactions in buffer X (30 mM tris-HCl pH 7.6, 60 mM KCl, 8 mM MgCl₂, 3 mM ATP, 5 mM 2-mercaptoethanol, 30 µg.mL bovine serum albumin) for lanes 1-8. Lanes 3-8 are serial dilutions of topo II: 1:2, 1:4, 1:8, 1:16, 1:32, and 1:64, respectively. Lanes 9-14 were reactions identical to those of lanes 3-8, except they were conducted in buffer X lacking ATP. Lanes 1 and 2 are supercoiled (SC) and relaxed (Rel) DNA markers, as indicated in the figure.
Table 1. Activity of topo II fractions through purification.

<table>
<thead>
<tr>
<th>sample</th>
<th>total protein (mg)</th>
<th>total activity$^a$</th>
<th>specific activity$^b$</th>
<th>yield</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3 M NaCl extract after Polymin P step.</td>
<td>16,250</td>
<td>ND$^c$</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>33-45% AS cut</td>
<td>360</td>
<td>1,600,000</td>
<td>4444</td>
<td>100%</td>
</tr>
<tr>
<td>BioRex 70 pool</td>
<td>6.4</td>
<td>640,000</td>
<td>110,000</td>
<td>40%</td>
</tr>
<tr>
<td>purified topo II</td>
<td>0.88</td>
<td>110,000</td>
<td>125,000</td>
<td>7%</td>
</tr>
</tbody>
</table>

$^a$Activity is expressed in units, where 1 unit is the amount of enzyme required to decatenate 250 ng kDNA in 15 min.
$^b$The specific activity is expressed in units enzyme/mg protein.
$^c$Activity can not be determined in the presence of Polymin P because this polycation interferes with the decatenation assay.
pool showed a significant increase in the specific activity of the topo II. The Phenyl-Sepharose column step resulted in a loss of total topo II activity, but is required to remove contaminating proteins, to concentrate the enzyme, and to remove the high salt concentration without a dialysis step (which results in a substantial loss of activity). As Table 1 indicates, the final yield of topo II in the PS pool is 7% of the activity in the AS cut, and contains about 0.9 mg total protein, for a specific activity of 125,000 units per mg protein. The purity of the topo II in the PS pool was examined on a Coomassie Blue stained SDS-polyacrylamide gel (Figure 5). The PS pool contains a single polypeptide of approximately 160,000 daltons, the same size determined for topo II purified from other organisms, suggesting that the PS pool is a substantially purified preparation of chicken topo II. Western blot analysis of the PS pool using a monoclonal antibody specific against MSB-1 topo II (a transformed chicken lymphoblastoid cell line) showed reactivity of only a single band, also at 160,000 daltons. Thus, the methods described above are a simple, rapid purification scheme for chicken erythrocyte topo II that yields a large quantity of essentially pure topo II (greater than 90% purity by protein gel electrophoresis). This purified topo II is stable indefinitely at -70°C (however, it is very susceptible to loss of activity by cycles of freezing and thawing), and is stable for weeks at -20°C
Figure 5. Topo II is a homogeneous band by SDS-polyacrylamide gel electrophoresis and by western blot analysis. Panel A shows a 12% SDS-polyacrylamide gel stained with coomassie blue. Lane 1 is molecular weight standards, as indicated, lane 2 is 100 μg of the topo II ammonium sulfate cut, and lane 3 is 1 μg of purified topo II. Panel B is a western blot of a gel as above. Lane 1 contained 40 μL purified topo II, lane 2 had 10 μL of purified topo II, and lane 3 contained 40 μL of topo II BioRex 70 pool. The proteins were transferred to nitrocellulose and incubated with a monoclonal antibody specific for topo II (provided by W. Earnshaw) and then with radiiodinated protein A, all by methods described in Materials and Methods. Protein size markers were run on the same gel then stained, and the molecular weights are indicated on the figure.
(where it does not freeze due to high ethylene glycol content) and for several days at 4°C.

Characterization

Characterization of the purified topo II showed ATP, MgCl₂, and cation requirements identical to those of topo II enzyme preparations from other organisms (data not shown). Additional experiments showed that magnesium, but not ATP, is required for formation of covalent topo II-DNA complexes, as determined by SDS/K+ precipitation (Figure 6A); however, topo II is less efficient than topo I for trapping of covalent intermediates. It was also found that certain metal ions inhibited the decatenation reaction catalyzed by topo II; an example is shown in Figure 6B, where 1 mM CuSO₄ was observed to completely block the reaction. I.K. Chung found that oligo dG inhibited DNA cleavage by topo II (unpublished results); oligo dG, but not oligo dC was also found to inhibit topo II decatenation of DNA at concentrations of 100 ng oligo per µL topo II or greater (Figure 6C), however a dimer of the human telomere repeat (TTAGGGTTAGGG) did not inhibit the enzyme (data not shown).

This simple and relatively rapid procedure for purification of topo II has also been applied to other cells. A similar, somewhat simplified, protocol was used to prepare fairly pure (about 50%) topo II from human placenta (Spitzner and Muller, 1988). The procedure has also been
Figure 6. Inhibition of topo II activity by several reagents. Panel A shows the results of an SDS/K+ assay titrating the amount of topo II added, in the presence or absence of MgCl₂. The procedure was as described in Materials and Methods. The SDS/K+ recoverable counts are shown relative to the amount precipitated by 5 μL topo I. Panels B and C show standard decatenation assays, with decatenated products labeled. Lane 1 is a linear kDNA marker, and kDNA was added to lanes 2-11; 2 μL topo II were added to lanes 4-11. Reactions were in topo II reaction buffer supplemented with CuSO₄ to the final concentration indicated: 8 mM in lanes 3 and 4, 4 mM in lane 5, 2 mM in lane 6, 1 mM in lane 7, 0.5 mM in lane 8, 0.25 mM in lane 9, 0.125 mM in lane 10, and 0.061 mM in lane 11. In Panel C, kDNA was added to all lanes and 1 μL topo II was added to lanes 2-7. Oligo dG [(dG)₁₂₋₁₈] was included in lanes 3-5, 30 ng, 100 ng, and 300 ng, respectively, and oligo dC [(dC)₁₉₋₂₄] was included in lanes 6 (100 ng) and 7 (300 ng).
adapted to purifying topo II from an MSB-1 cell nuclear extract, from *Saccharomyces cerevisiae*, and from *Drosophila melanogaster* (data not shown). It was also found that increasing the 2-mercaptoethanol concentration during the extraction from chicken erythrocyte nuclei from 30 mM to 140 mM greatly increased the yield and specific activity of the topo II preparation. As Figure 7 shows, topo II purified in this fashion showed an activity of approximately 100 units per μL (which can be increased somewhat by concentration with a Micro-Pro-Dicon unit), and a specific activity of about 2,000,000 units per mg protein. However, this extraction procedure leads to a preparation in which minor contaminants were observed on SDS-polyacrylamide gels, so the use of an increased amount of 2-mercaptoethanol has not been continued, as the homogeneous topo II preparation is more desirable for further studies, to prove that the enzyme is responsible for whatever activities are observed.
Figure 7. Titration of high activity topo II preparation. Lane 1 is a linear kDNA marker, lane 2 is kDNA, lane 13 is a topo II marker (products are indicated as OC and Rel), and lanes 2-12 are standard decatenation reactions assaying serial dilutions of topo II prepared from 2-mercaptoethanol extracted cells (lanes 4-7) and the same material after concentration (lanes 8-12). Dilutions were 1:3 (lanes 4 and 8), 1:9 (lanes 5 and 9), 1:27 (lanes 6 and 10), 1:81 (lanes 7 and 11) and 1:243 (lane 12).
DISCUSSION

Existing protocols for purification of topo II required numerous steps, so a more simple and rapid procedure was desired for purifying chicken erythrocyte topo II. It had been shown that chicken erythrocyte nuclear extracts contained topo II (not shown), and that while 0.4 M NaCl was required for efficient extraction of topo I (Trask and Muller, 1983), 0.3 M NaCl was sufficient for extraction of topo II with fewer contaminants. 2-mercaptoethanol was essential for good yields of topo II (presumably to release it from the nuclear matrix; see Earnshaw and Heck, 1985); however, inclusion of high concentrations resulted in maximal topo II activity (see Figure 7), but at the expense of purity. Polyethyleneimine was useful for removal of nucleic acids concurrent with the precipitation by centrifugation of nuclear debris; under the extraction conditions, topo II remained in the supernatant. In a buffer of lower ionic strength, Goto et al (1984) found that topo II precipitated with this reagent, and they used this as a part of their lengthy purification scheme.
Ammonium sulfate (AS) fractionations were used both to partition topo II away from the polyethyleneimine in the solution that interferes with decatenation of kDNA by topo II (data not shown) and for further purification; only about 2% of the protein from the nuclear extract was precipitated in the 33-45% AS cut that contained the topo II activity. The AS cut was diluted to about 0.2 M salt and loaded onto a BioRex 70 column. This cation exchange matrix was selected because of its high protein binding capacity and the rapid flow rate permitted. Gradient fractions were assayed by decatenation of kDNA which is rapid and specific for topo II (Marini et al, 1980); the enzyme eluted in a broad peak from about 0.35 to 0.45 M potassium phosphate (see Figure 2). The BioRex pool had a very low protein concentration, but resulted in a 25-fold purification of topo II activity compared to the AS cut (Table 1).

The BioRex pool was loaded directly onto a Phenyl-Sepharose column, which was subsequently washed with a buffer containing 20% ethylene glycol and very little salt; under these conditions most proteins wash right through this hydrophobic affinity resin (see Trask and Muller, 1983), but topo II remained bound. Washing of the column with a buffer containing 60% ethylene glycol was sufficient to elute topo II activity in a concentrated pool. The yield off this column was not high (usually about 1 mg; see Table 1), but the eluted material was shown to be pure topo II (see the
picture of the SDS-polyacrylamide gel in Figure 5A) with a molecular weight of approximately 160,000 dal. In addition, the Phenyl-Sepharose pool did not require dialysis, because there was little salt to interfere with ensuing reactions, and the ethylene glycol appeared to aid in stabilization of the enzyme. This pool retained activity indefinitely at -70°C (at least 3 years) and was stable for several months at -20°C and several days at 4°C.

The topo II purified from chicken erythrocytes cross-reacted with an antibody generated against MSB-1 topo II (Figure 5B) and exhibited properties similar or identical to other eukaryotic type II topos that have been characterized. The novel purification scheme has also been shown to be useful for isolating topo II from other sources; however, the eluant from the second column is not always purified to homogeneity (see Spitzner and Muller, 1988).
CHAPTER II

Generation of monoclonal antibodies against chicken topoisomerase II.

INTRODUCTION

Specific antibodies are valuable reagents for the study of proteins. In particular, they can be used to quantitate levels of protein in cell extracts, and can be utilized to trace a protein through a purification procedure when other means of assay are difficult (such as conditions under which an enzyme is inactive, or other factors have competing activities). Antibodies are also of use for immunoprecipitation and for cytological localization of an antigen. Immunofluorescent microscopy was used by Earnshaw and Heck (1985) to identify topo II as Sc-1, a major component of the nuclear scaffold.

However, antibodies were not readily available, and the cross reactivity of other antibodies against chicken erythrocyte topo II was unknown. Therefore, it was decided that antibodies would be generated. Monoclonal antibodies were made because they are more specific (less background binding than a polyclonal antibody), and can be selected to recognize desired epitopes of an antigen. A mouse was
immunized with purified chicken topo II, and hybridomas were generated and screened for binding of topo II. Several lines of antibodies were identified that showed specific recognition of topo II on western blots, one of which also bound human topo II. These antibodies were found not to bind topo II under conditions for immunofluorescent microscopy, however.
MATERIALS and METHODS

Materials. Chemical reagents were purchased from Sigma. Immunological reagents were from Cappel. Vectastain kit was from Vector Laboratories. Radioiodinated Protein A was from ICN. Tissue culture media were from Gibco and MA. Tissue culture vessels were from Corning and Falcon. A1 myeloma cells and other reagents for generating the hybridomas were kindly provided by D. Galloway.

Methods. Macrophage feeder layers were made by killing balb/c mice by cervical dislocation, cutting the skin away, and injecting the peritoneum with 5 mL sterile 0.34 M sucrose; after massaging the region near injection, the fluid was collected using a 10 mL syringe. Macrophages were washed, then resuspended in RPMI medium at 200,000 cells per mL. One drop was added to each well of 96 well plates, and 1 mL per well of 24 well plates.

ELISA assays were conducting using the Vectastain kit and procedures. Topoisomerase II was coated in wells at 200 ng (2 μL enzyme) per well in 0.1 M sodium carbonate (pH to 9.6
with HCl). The wells were washed with PBS (10 mM sodium phosphate pH 7.5, 145 mM sodium chloride), then the primary antibody was incubated in the wells for several hours at 37°C. The Vectastain procedure uses a biotinylated anti-mouse secondary antibody, followed by an avidin reagent conjugated to alkaline phosphatase. The substrate for the alkaline phosphatase assay was p-nitrophenyl phosphate (tablets from Sigma). Antibody typing was performed using a Boehringer Mannheim Biochemicals kit as recommended. Protein gel electrophoresis and western blot procedures were as described in Chapter 1. The MSB-1 cell extract was prepared simply by boiling a cell pellet in the denaturing/reducing loading dye.
RESULTS

Generation of monoclonal antibodies

One balb/c mouse was injected intraperitoneally with 50 μg purified topoisomerase II in complete Freund's adjuvant. After three weeks, and again three weeks after that, the mouse was boosted with an additional 50 μg antigen. The tail was nicked to get blood, which was tested by ELISA against purified topo II, to determine that the mouse was producing antibody against topo II. Three weeks after the second boost, the mouse was killed by cervical dislocation, and the spleen was aseptically removed. The spleen was broken open and the spleen cells were separated from one another by gently forcing a metal fine mesh screen over the spleen. After additional pipeting to separate cells from large fragments of the capsule, the cells were transferred to a tube. Clumps of cells were allowed to settle out, then the supernatant was transferred to another tube, and the spleen cells were collected by centrifugation. Cells were resuspended in sterile, buffered saline at 10^8 cells per mL. A1 myeloma cells were grown in RPMI supplemented with 10% fetal bovine serum, L-glutamine, sodium pyruvate, and
gentamycin (all as recommended 1x concentrations), then washed and resuspended as were the spleen cells, also at $10^8$ cells per mL. One mL each of spleen cells and myeloma cells were mixed, centrifuged, and resuspended in 1 mL 50% PEG [5 g polyethylene glycol 4000 autoclaved then mixed with 5 mL 10% DMSO (aqueous)], then the tube was slowly rotated for 1.5 min at 37°C. Then buffered saline was added slowly up to 50 mL, then the cells were pelleted and resuspended in RPMI medium with HAT (13.6 μg/mL hypoxanthine, 0.176 ng/mL aminopterin, 3.71 μg/mL thymidine), and the fusion mixture was added to five 96 well plates preset with macrophage monolayers at one drop per well. Plates were incubated in a humidified incubator in 95% air/5% CO2 at 37°C; after 4 and 8 days, 1 drop fresh RPMI-HAT was added to each well.

**Screening of monoclonal antibodies**

After colonies began to appear, hybridoma supernatants were screened by ELISA for specific binding to topo II. Clones that gave positive signals were expanded into 24 well plates (with macrophage feeder layers, in RPMI-HAT) and rescreened. Clones that gave the strongest ELISA signals were then cloned by limiting dilution in 96 well plates, and grown as before. There were three cell lines that were recloned, called B1, E1, and B5. For each of these, a number of clones were selected that gave strong ELISA
signals; these were expanded up to growth on 100 mm petri dishes and transferred to growth in just RPMI with 10% fetal bovine serum and gentamycin (50 μg/mL), without macrophage feeder layers. These hybridoma cell lines were carried as suspension cultures in 60 and 100 mm petri dishes and were split 1:10 once every 10-15 days. The cell lines were also frozen in liquid nitrogen.

**Antibodies react specifically with topo II**

The hybridoma lines continued to react with purified topo II as measured by ELISA over many months. To examine the specificity against topo II, western blots were run. In the first experiment, shown in Figure 8, an MSB-1 cell extract (a transformed chicken lymphocyte cell line) and purified topo II were electrophoresed on a 7% SDS-polyacrylamide minigel and transferred to a nitrocellulose membrane. The filter strips were incubated with the different primary antibodies (including several of the hybridoma supernatants and a known monoclonal antibody to chicken topo II, prepared against the protein in the chicken tumor cell line MSB-1 by W. Earnshaw [Earnshaw and Heck, 1985]). The Vectastain kit was used as described by Vector Laboratories (see Materials and Methods) to detect mouse antibodies that had bound to proteins on the nitrocellulose membrane. As the results illustrate, all the tested
Figure 8. Preliminary western blot analysis of antibodies using a whole cell MSB-1 extract. An SDS-polyacrylamide gel was run with purified topo II (30 μL, labeled "T") and a whole cell MSB-1 extract (30 μL, labeled "M"). Proteins were transferred to nitrocellulose and incubated with the antibodies, then the blot was developed using the Vectastain kit and the phosphatase substrate. The antibody samples were (as labeled above the gel): 1. a known monoclonal antibody directed against MSB-1 cells (see Earnshaw and Heck, 1985), 2. E1 hybridoma supernatant, 3. serum taken from the topo II-immunized mouse before spleen removal, 4. B1 hybridoma supernatant, 5. B5 hybridoma supernatant, and 6. another E1 hybridoma supernatant. The three major bands that show binding of the antibodies to the MSB-1 cell extract are labeled A, B, and C in the figure. A is the expected intact topo II molecule, while B and C are probably proteolytic fragments.
reagents bound essentially the same polypeptides. The three bands of greatest intensity for the hybridoma supernatants tested against the MSB-1 extract appear to line up exactly with the stained bands observed with the Earnshaw topo II antibody; similarities were also observed between the two lanes with purified topo II. These results suggested that the hybridomas might be producing anti-topo II antibodies.

The hybridoma cell lines were recloned to insure that no contaminating antibodies were being produced. The three hybridoma lines were compared with the monospecific Earnshaw topo II antibody for binding to the whole cell MSB-1 extract (M) and to purified topo II (T); the results are shown in Figure 9. Antibodies 1, 2, 3, and 4, are, respectively, B1 supernatant, Earnshaw monoclonal, E1 supernatant, and B5 supernatant. In the cell extract lane, all 4 antibodies light up very strongly a 160,000 dal protein that is presumably topo II, and a few weaker subbands. Most of the subbands in the hybridoma supernatant lanes correspond exactly to bands which appear with the Earnshaw antibody as well, and the B1 supernatant bands show the same bands and intensities as the Earnshaw antibody. All the antibodies bound a major band at 160,000 in the purified topo II lanes, and the minor subbands are also the same for all antibodies. After further subcloning over a number of months, the different lines were tested by western blot analysis with an MSB-1 whole cell extract to see which were
Figure 9. Western blot analysis of hybridomas against a whole cell extract and against purified topo II. Protein samples were MSB-1 whole cell extract ("M") or purified chicken topo II ("T"); antibodies were analyzed as described above. Antibodies tested were: 1, B1; 2, Earnshaw; 3, E1; and 4, B5. The expected mobility of topo II (160,000 dal) was deduced from molecular weight markers (not shown) and is indicated on the figure.
the most specific against topo II. In Figure 10, instead of using the Vectastain immunoassay kit, after incubation with primary antibodies, the membrane was incubated with a rabbit anti-mouse antibody, then with radiiodinated Protein A and autoradiography. The secondary antibody was used because neither the Earnshaw antibody nor any of the monoclonal lines tested binds Protein A directly. In Figure 10, two sets of bands can be seen; a high molecular weight one (about 160,000 dal from molecular weight standards that were run on the same gel, not shown) and one at about 66,000 dal. The lower (stronger) band is nonspecific; it appears in all experiments using this cell extract and the Protein A. Lane 9 was not incubated with any primary antibody, yet that band remains (and if secondary antibody is deleted, the band is still there, S. Ebert unpublished result). It is therefore the upper band that is of interest. The Earnshaw antibody (lanes 1 and 15) yields a clean band. Several hybridomas showed this same binding (lanes 6, 11, 14, and possibly 4 and 10). In subsequent western blot analysis experiments, these same hybridoma lines continued to exhibit binding behavior identical to that of the Earnshaw topo II monoclonal antibody. In addition, one line, B5-B3, was found to also bind specifically to human topo II in whole cell extracts or with purified enzyme (S. Ebert, unpublished results); the Earnshaw antibody does not cross-react with human topo II.
Figure 10. Western blot analysis of anti-topo II antibodies using radiiodinated Protein A. An MSB-1 whole cell extract was run on an SDS-PAGE gel as one long lane; after transferring the proteins to nitrocellulose, the membrane was cut into strips and incubated with the following antibodies: Lanes 1 and 15, Earnshaw topo II antibody; lane 2, purified B1-D7 antibody (a B1 subclone); lane 3, purified E1-B12 (an E1 type); lane 4, E1-H7 ascites fluid (an E1 type); lanes 5-8 B5-A1, B5-B3, B5-C5, and B5-A1 (all B5 subclones); lane 9, no primary antibody; lanes 10-12, B1-H10, B1-H2, and B1-D7 (B1 clones); and lane 14, E1-H7 supernatant. The intense lower band is nonspecific, as it also appeared in the absence of primary antibody (lane 9). The upper band is the topo II-specific band (compare to the control antibody in lanes 1 and 15).
Several of the monoclonal antibody lines have been expanded in vivo. Mice were injected intraperitoneally with 0.5 mL Pristane (2,6,10,14-tetramethylpentadecane), then three weeks later 0.5-1.0 x 10^6 hybridoma cells were injected. Mice that showed abdominal swelling after several weeks were punctured with needles and ascitic fluid was collected. The ascites fluids showed protein binding behavior identical to that of the supernatants from which they originated. Some immunoglobulins were purified by Protein A-Sepharose column chromatography (by S. Soisson, unpublished result) and these also behaved as the supernatants.

**Characterization of antibodies**

Antibody typing kits were used to determine the type of immunoglobulin for each of the hybridomas. All lines had Kappa light chains. The B1 line was clearly of the IgG2B type light chain (this is the same subtype as the Earnshaw antibody). The E1 lines are also IgG, but the results were inconclusive (ELISA showed equal responses with IgG1 and IgG2B subtypes). In addition, the B1 and E1 lines demonstrated mobilities of IgG and not IgM chains after analysis by SDS-polyacrylamide gel electrophoresis, followed by staining with Coomassie Blue. The B5-B3 monoclonal antibody (the one that appears the most specific and the
most useful because it cross-reacts with human topo II) has failed to react well with any of the subtyping reagents. Preliminary results suggest that it may be either an IgM or an IgG type. The fact that the antibody does not bind Protein A strongly suggests that it is not a common IgG type. Further tests will elucidate the characteristics of the monoclonal antibodies.

The antibodies do not inhibit topoisomerase II activity. Incubation of purified topo II with these antibodies (90 min preincubation at 4°C) does not inhibit decatenation or relaxation activities at all (data not shown). Furthermore, in an experiment in which topo II-DNA covalent complexes were incubated with one of the antibodies, then incubated with Protein A-Sepharose beads, radiolabeled DNA was precipitated with the beads in proportion to the amount of topo II added (in other words, the more topo II added to the indirect immunoprecipitation experiment, the more DNA was recovered with the beads, data not shown). In addition, none of the hybridomas tested has been useful for immunofluorescent microscopy (in direct comparison, the Earnshaw antibody exhibited intense nuclear staining when chicken embryo fibroblast cells were studied). Together, these results suggest that the hybridomas are specific antibodies against topo II, but are not directed against the same epitopes of the enzyme as the Earnshaw antibody.
DISCUSSION

Characterization of monoclonal antibodies against topo II

A mouse was immunized with chicken topo II, then its spleen cells were fused with myeloma cells. These hybridomas were screened and colonies were selected that bound topo II (as assayed by ELISA). The monoclonal antibodies were shown to be specific against topo II by western blot analysis (see Figures 8-10) and to exhibit banding patterns identical to that of a known monoclonal antibody against topo II (Earnshaw and Heck, 1985), with the major band at 160,000 dal (the size of topo II). The other immunoreactive bands observed are presumably due to proteolytic degradation of topo II. One piece of evidence supporting this idea is that as purified preparations or crude extracts with topo II are thawed after being refrozen, smaller bands of greater number and intensity than when freshly prepared are observed on western blots (data not shown). Additionally, the fact that the lines are recloned monoclonals would suggest that only a single antibody species is produced, and subbands with purified topo II must be related to topo II (proteolytic fragments). The differences in reactivity (the pattern of bands) implies
that different antibody lines recognize different epitopes on the topo II molecule.

One antibody line also cross-reacted with human topo II (unlike the antibody of Earnshaw, data not shown). None of the antibodies bound to $^{125}$I-Protein A directly (secondary antibodies had to be used for the western blots). In addition, the antibodies did not inhibit the decatenation of kDNA by topo II, and none of the lines was useful for immunofluorescent microscopy (in contrast to the Earnshaw antibody). The results suggest that the antibodies bind to epitopes of topo II that are not exposed when the enzyme is a part of the nuclear matrix (such as at sites for protein-protein interaction). These antibodies have proven to be of use for quantitating topo II levels in samples. They were used to measure the amount of topo II precipitated in enzyme-oligo dG aggregates (I.K. Chung and M.T. Muller, submitted), and were used by S. Ebert to examine the effect of HSV-1 infections on topo II levels in cell extracts (unpublished data). Additional experiments will be conducted to determine whether the antibodies will be of value for immunofluorescent microscopy to localize the enzyme within cells during HSV-1 infection; new fixatives will be used in place of formaldehyde (which was used previously, without success).
Chapter III

Derivation of a consensus sequence for DNA cleavage by topoisomerase II.

INTRODUCTION

An important objective in understanding the functions of topo II was to characterize the DNA sequences that it interacted with. The sites of topo II recognition can be determined by trapping the covalent protein-DNA intermediate and sequencing the DNA break (see reviews by Wang, 1985; Vosberg, 1985). This makes it much easier to map topo II than DNA binding proteins that do not cleave DNA. It was shown that topo II cleaved DNA nonrandomly, but topo II bound DNA much more promiscuously than most site-specific binding proteins. Thus it was originally suggested that topo II sites could not be specified by a consensus sequence (see Wang, 1985). However, it was more likely that DNA sequence did specify topo cleavage sites, but the recognition elements would be highly degenerate. For this reason, it was decided that in order to generate a useful consensus sequence, a very large number of sites should be
sequenced, from substrates with a wide range of base compositions. A consensus sequence was determined for Drosophila topo II, but Sander and Hsieh (1985) analyzed only 16 sites, all from a single DNA sequence; therefore, the generality of their sequence was suspect. Chicken topo II cleavages were sequenced for 71 strong sites and a consensus sequence was derived and shown to be of value for predicting topo II sites. Other analyses confirmed that the recognition elements were within the window suggested by the consensus sequence (directly flanking the cleavage site). For example, an analysis of sequence redundancy by position (a measure related to the information content of a data base; see Stormo, 1988) indicated that the nucleotide distributions in the 71 sequences were essentially random except at the consensus sequence positions.
MATERIALS AND METHODS

Materials. Restriction enzymes were purchased from Bethesda Research Laboratories, T4 polynucleotide kinase was from United States Biochemical Corporation (gamma-$^{32}$P)-ATP was from ICN, and chemical reagents were from Sigma.

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 β-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 $10 	imes 10^4$ 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: 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 MgCl₂ 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 10⁴ 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 ladders 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 arbitrarily 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 2).

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 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 4 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 4 consensus were evaluated. This consensus represents the highest ratio of real data match: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 increase 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.
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 (related to information content, see Stromo, 1988) is derived mathematically as follows.

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

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

\[ (2) \quad (4)(-0.25)(-2) = 2.0 \]

The percent redundancy therefore is:

\[ (3) \quad 1 - \frac{\text{actual } U(x)}{\text{maximum } U(x)} \times 100\% \]

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 12. If nucleotide proportions at a position are essentially random, then redundancy approaches zero (see Figure 12).
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 2). Reactions were carried out in the presence and absence of the topoisomerase II inhibitor m-AMSA and were analyzed on sequencing gels with chemical sequence ladders (see Spitzner and Muller, 1988) to resolve cleavages to single residues. When possible, both top and bottom strands were analyzed (see Table 2). 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
Table 2. DNA substrates used to derive consensus sequence

<table>
<thead>
<tr>
<th>Sequence Number:</th>
<th>DNA Source:</th>
<th>Restriction sites (size [position]):</th>
<th>Comments on the sequence (ref):</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>kDNA</td>
<td>*Stul-HinfI (226 bp) [2515-2289]</td>
<td>Bent DNA fragment (see footnote c for source of sequence)</td>
</tr>
<tr>
<td>2.</td>
<td></td>
<td>*Stul-HinfI (549) [1-549]</td>
<td>footnote c</td>
</tr>
<tr>
<td>3.</td>
<td></td>
<td>*Avall-SacII (258) [1763-2020]</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td></td>
<td>*Hinf-SacII (263) [2283-2021]</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td></td>
<td>*Xhol-HinfI (723) [1272-549]</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td></td>
<td>*Xhol-HinfI (1011) [1273-2283]</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td></td>
<td>*Hinf-HaeIII (421) [549-129]</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>Yeast mitochondria</td>
<td>*HpaI-EcoRI (345)</td>
<td>S. cerevisiae (strain SD23) Var1 gene coding region (26)</td>
</tr>
<tr>
<td>9.</td>
<td></td>
<td>*EcoRI-HpaI (345)</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>pUC12</td>
<td>*PvvI-EcoRI (90) [306-396]</td>
<td>lac z gene region (27)</td>
</tr>
<tr>
<td>11.</td>
<td>archaeobacteria</td>
<td>*HpaI-EcoRI (790) [+592/-101]</td>
<td>5' region of methyl reductase gene of M. vanneili (28)</td>
</tr>
<tr>
<td>12.</td>
<td>HSV-1</td>
<td>*HindIII-EcoRI (352)</td>
<td>Ori fragment in pON103 (29)</td>
</tr>
<tr>
<td>14.</td>
<td></td>
<td>*NcoI-Aval (121) [-204/-325]</td>
<td></td>
</tr>
<tr>
<td>15.</td>
<td></td>
<td>*Aval-NcoI (192) [-12/-204]</td>
<td></td>
</tr>
<tr>
<td>16.</td>
<td>HSV-1</td>
<td>*Aval-NcoI (1211-325/-204)</td>
<td>as above in fragment #15</td>
</tr>
</tbody>
</table>
Table 2 (cont.)

| 17. | *HindIII-PvuII | glycoprotein D promoter region (31) |
| 18. | X. laevis | *HindIII-Rsal (111) [560-670] 5S RNA gene 5' flank and coding regions (32) |
| 19. | *Rsal-HindIII (111) [670-570] |
| 20. | Human DNA | *Acll-Sphl (245) [61816-61571] 5'-globin gene, 5' region, 52 bp purine-pyrimidine Repeat (Genbank) |
| 21. | *Hpal-Acll (444) [61372-61816] 5'-globin 5' region (Genbank) |
| 22. | *BamHI-EcoRI (918) [62666-63582] 5'-globin coding region (Genbank) |
| 23. | *EcoRI-BamHI (918) [63582-62666] |
| 24. | bacteriophage lambda | *EcoRI-BglII (359) [39168-38814] Origin of replication (Genbank) |
| 25. | SV40 | *Neol-Rsal (258) [37-295] Regulatory region (Genbank) |
| 26. | SV40 | *Neol-AvalII (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, orientation 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.
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 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 (reviewed by Vosberg, 1985 and Wang, 1985), which effectively changes the position of cleavage relative to nucleotide positions as assigned in the data base. The need for alignment of sequences before a consensus sequence can be found is illustrated in Table 3. In Panel A, the sequences are shown with the top strand as the strand that was observed from the analysis of the cleavage site on the original sequencing gel. In the same manner as with the actual topo II cleavage data, the sites as observed show very little homology to one another, and the only "consensus sequence" that can be deduced contains little information. However, if the orientation of some of the sequences is reversed, so that what was the unseen bottom strand is now the top strand, as shown in Panel B, a much more specific and useful consensus sequence can be deduced, and it can be seen that the sequences demonstrate a 69% average homology to this consensus sequence that could not have been
Table 3. Consensus sequence information must be recovered by alignment.

<table>
<thead>
<tr>
<th>Seq#</th>
<th>DNA Sequence</th>
<th>Match</th>
<th>Consensus Sequencea</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>*GG^ TT TT AT</td>
<td>67%</td>
<td>NN ^ NN NT NC</td>
</tr>
<tr>
<td>1</td>
<td>CC A AAAA ^ TA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>*TT^ GC AT CC</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AA CG TA ^ GG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>*AT^ TAGT TA</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TA AT CA ^ AT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>*TC^ CG AT TT</td>
<td>67%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AG GCT A A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>*AC^ ACC C AC</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TG TG GG ^ TG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>*CA^ AAC G GA</td>
<td>67%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GT TT GC ^ CT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>*GC^ TCCA AC</td>
<td>83%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CG AG GT ^ TG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>*GG^ GAT C GC</td>
<td>67%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CC CT AG ^ CG</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)Bases were assigned to the consensus sequence if the proportion of a single base at a position was at least 1/2, or, failing this, if 2 bases together comprised at least 2/3 of the base occurrences at a position. The match to the consensus sequence is the percentage of the non "N" bases underlined in the consensus sequence above.
determined without some way of choosing how to orient the sequences before knowing what to orient them with respect to. This, then, was a major obstacle in determining a consensus sequence for topo II, because the sequences did not have much in common. The problem was tackled by use of an iterative alignment method. In the original data base, the base proportions were essentially random, but at a few positions one base was observed to have slightly higher frequencies of occurrence, notably a +4 T, a +2 G, and a -2 G (not a useful consensus sequence, however). These bases were used for the initial alignment of sequences. A computer program was designed to determine which strand of a topo II cleavage site showed a better match to the input "consensus sequence" and then oriented each site so that the top strand was now the strand with the better match. Then the base proportions at each position relative to the cleavage site were determined again, a new "consensus sequence" was derived based on which bases now stood out, and this new sequence was used as the template to realign the sites. After each repetition, new "consensus sequence" elements were observed. Finally, continued iterations produced no new alignment of the sites; the set of sequences was then locked in to this new alignment for further analysis. The distributions of some nucleotides relative to the cleavage site are graphed in Figure 11; the proportions deviate significantly at some positions proximal to the
Figure 11. Analysis of nucleotide frequencies. In figures A through D, the proportion or frequency of occurrence of purines (panel A), thymine (panel B), purine-N-pyrimidine (panel C), and the trinucleotide GGT (panel D) are shown at each position relative to the cleavage site (marked on the X axis). The data were compiled from the data base of 71 strong topo II sites from Table 2. 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.
cleavage site and are essentially random at positions outside a window flanking the cleavage site.

Redundancy analysis is a measure of the deviation from random base proportions at a position. This measure is useful for suggesting which positions contain base information at greater than background frequencies, in other words, which positions are candidates for inclusion in a consensus sequence. The percent redundancy at each position is shown graphically in Figure 12; a redundancy percentage near zero indicates random base proportions at that position (so that it would be represented as an N, for any base, in a consensus sequence). The positions that stood out by this analysis are, relative to the cleavage site, -10, -8, -5, -2, -1, +2, +3, +4, +6, +8, +10, and +11. The "rules" for inclusion of a base in a consensus sequence are somewhat arbitrary; the elements selected for the topo II consensus sequence were those with greater than 50% proportion of a single base at a position, or with greater than 70% proportion of two bases combined at a single position. However, since a consensus sequence has, at best, predictive rather than simply descriptive value, positions +10 and +11 were excluded from the consensus sequence because the inclusion of pyrimidines in a "consensus sequence" at those positions actually lowered the average match of the cleavage sites to that "consensus", while increasing the match of a set of randomly generated sequences to that "consensus
Figure 12. Analysis of redundancy in topo data base. The percentage redundancy was plotted (see Materials and Methods) 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 non-random distribution of bases. Peaks are seen only at the non-"N" consensus positions except for +10 and +11 (see text). All other positions show a redundancy near zero, indicating random distributions at these positions.
sequence". Therefore, the consensus sequence determined for topo II contains the elements which give the best discrimination between topo II sites and nonsites. The topo II consensus (including the base proportions used) is shown in Table 4. This consensus sequence is an 18 base sequence flanking the cleavage site, of which 10 positions are specified (or non-"N" bases). The 71 sites in the data base have an average match to the 10 non-"N" positions of 69.6%, or about 7 out of 10 (and random bases match about 4 out of 10). This analysis suggests that topo II makes its base-specific interactions with DNA in the 18 base consensus region proximal to the site of cleavage; however, this does not preclude protein-DNA interactions in flanking regions. Indeed, DNase I footprinting experiments suggest that topo II protects a wider region than the 18-base consensus sequence (Lee et al, 1989).

**Consensus is the major determinant for DNA cleavage by topo 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;
Table 4. Nucleotide frequencies used to derive the topoisomerase II consensus

<table>
<thead>
<tr>
<th>-10</th>
<th>-9</th>
<th>-8</th>
<th>-7</th>
<th>-6</th>
<th>-5</th>
<th>-4</th>
<th>-3</th>
<th>-2</th>
<th>-1</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.75</td>
<td>0.62</td>
<td>0.22</td>
<td>0.87</td>
<td>0.14</td>
<td>0.52</td>
<td>0.33</td>
<td>0.19</td>
<td>0.56</td>
<td>0.28</td>
<td>0.17</td>
<td>0.12</td>
<td>0.26</td>
<td>0.23</td>
<td>0.06</td>
<td>0.12</td>
<td>0.31</td>
<td>0.13</td>
</tr>
<tr>
<td>0.19</td>
<td>0.14</td>
<td>0.23</td>
<td>0.32</td>
<td>0.11</td>
<td>0.13</td>
<td>0.16</td>
<td>0.11</td>
<td>0.11</td>
<td>0.13</td>
<td>0.15</td>
<td>0.11</td>
<td>0.21</td>
<td>0.21</td>
<td>0.21</td>
<td>0.33</td>
<td>0.33</td>
<td>0.07</td>
</tr>
<tr>
<td>0.12</td>
<td>0.13</td>
<td>0.21</td>
<td>0.33</td>
<td>0.18</td>
<td>0.11</td>
<td>0.16</td>
<td>0.11</td>
<td>0.11</td>
<td>0.13</td>
<td>0.15</td>
<td>0.11</td>
<td>0.21</td>
<td>0.21</td>
<td>0.21</td>
<td>0.33</td>
<td>0.33</td>
<td>0.07</td>
</tr>
<tr>
<td>0.12</td>
<td>0.13</td>
<td>0.21</td>
<td>0.33</td>
<td>0.18</td>
<td>0.11</td>
<td>0.16</td>
<td>0.11</td>
<td>0.11</td>
<td>0.13</td>
<td>0.15</td>
<td>0.11</td>
<td>0.21</td>
<td>0.21</td>
<td>0.21</td>
<td>0.33</td>
<td>0.33</td>
<td>0.07</td>
</tr>
</tbody>
</table>

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'
thus, GC and GT were the most common dinucleotides found at positions -2, -1, but other bases were also found. A search for palindromes was performed to investigate whether topoisomerase II recognizes any 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 sequences for each DNA 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. Our conclusion is that topoisomerase II does not appear to require any dyad axis of symmetry in strong cleavage sites, and this is clearly reflected in the consensus sequence. We additionally performed several analyses on the data base designed to uncover subtle features in recognition that might be missed by a linear search of sequence repeats and frequencies. A Fourier analysis was performed to determine whether there were relevant periodicities within sequences cleaved by topoisomerase II. We did not detect any significant underlying periodicity in the nucleotide occurrence. We performed Markov analysis to examine whether a base at a given position affected the probability of
finding a specified base at another position. For example, when we tested the dependence of a consensus match on whether the previous consensus position matched, the Chi square values were not a significant. From these data we conclude that the probabilities of base occurrence are independent of one another, and its 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, as in this data base), the expected frequency of matching a consensus sequence depends solely on the number of positions in the consensus sequence, the degree of specificity at each position, and the matching criteria selected. For example, a 70% match to the consensus shown in Table 4 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 is about one per 25 bases (strong + weak sites), a consensus of the size and degeneracy shown in Table 4 is consistent with the observed frequencies of topo II sites. Similarly, the strongest topo II sites are observed at a frequency of 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 although in reality the consensus contains 18 bases. Topoisomerase II cleavages were performed and the consensus homologies at different sites compared to actual cleavage data. As shown in Figure 13, 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 (see reviews above), its electrophoretic mobility is somewhat different than the chemical sequence marker which contains a 3'-phosphoryl end (see also Muller et al, 1988). 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
Figure 13. Topoisomerase II cleavage on HSV-1 promoter fragment. Cleavage reactions were carried out on fragment number 14 (top gel) or number 16 (bottom gel) from Table 2. 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 topo II plus 50 μg/mL m-AMSA; lanes 4 and 9 contained 4 units topo II; and lanes 5 and 10 received no enzyme. Above each gel is shown the actual sequence along with a histogram showing percentage match to consensus at each position. Each base in the sequence was treated as +1 in the consensus (i.e. 1 base 3' of the cleavage site) with the flanking bases as -10 to +8. The consensus was moved as a window along the portion of the sequence that was resolved on the gel. Only matches to non-"N" sites were 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. The arrows designating sites A through E correspond to the topo II sites discussed in the text.
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 13 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 hr exposure) had an average match of 72% on the "best matched" strand and on the other strand the average match was 45%. Moderately strong cleavages (clearly visible bands after 24
hr exposure) matched an average of 58% (best matched strand) and 42% on the opposite strand, while weak cleavages (barely detectable bands in 24 hr 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 is 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 14A shows the nucleotide frequencies at each position relative to the cleavage site from the data base of 142 sites. Figure 14B depicts the redundancy at each position. Clearly, the symmetric data still contain significant sequence information, as seen by the peaks (Figure 14B) which were used to define consensus elements. The resulting symmetric consensus is more degenerate than the top strand (or asymmetric) consensus shown in Figure 14. Superficially, the symmetric consensus derived from the 142
Figure 14 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 likely define the consensus sequence.
sites resembles a blend of the top strand and the bottom strand (taking into account the 4 base 5' 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 cleavages with reasonable accuracy, but the symmetric consensus sequence was prone to predict more false positives.

One way to express the information present on both strands of a cleavage site is to determine the number of matches to the asymmetric consensus sequence on the top strand and then add to this the number of matches to the same asymmetric consensus sequence (not its complement) on the bottom strand; this measure is called the combined score (or combined match). In an example (see Chapter V) testing the value of the combined score for predicting topo II cleavage sites, the threshold value to maximize accuracy for the combined score was selected as 10; using this score, the majority of cleavage sites were predicted with few false positives; strong cleavage sites were predicted with much greater accuracy than weaker sites.
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 (Muller and Mehta, 1988; see references in the Introduction). 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 (see above references) and sites are enriched in intergenic regions. Clustered in vitro cleavages by topoisomerase II were also reported in an enhancer region of c-fos (Darby et al, 1986), 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 (Rowe et al, 1985; Uvdardy et al, 1985; Muller and Mehta, 1988) and in SV40 where the strongest in vivo site (near nucleotide 270) is present in the control region (Yang et al, 1985). 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 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 (Yang et al, 1985). 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-IID 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-IID factor binding are underscored):

<table>
<thead>
<tr>
<th>base:</th>
<th>270</th>
<th>271</th>
<th>272</th>
<th>273</th>
<th>274</th>
<th>275</th>
<th>276</th>
<th>277</th>
<th>278</th>
<th>279</th>
<th>280</th>
</tr>
</thead>
<tbody>
<tr>
<td>SV40:</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>T</td>
<td>G</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>Topo:</td>
<td>N</td>
<td>N</td>
<td>G</td>
<td>Y</td>
<td>N</td>
<td>G</td>
<td>G/T</td>
<td>T</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
</tbody>
</table>

Note that methylation interference is seen at conserved G residues that match the topoisomerase II consensus. Furthermore, the GT-IID complexes cannot be effectively competed by fragments that do not match our consensus sequence but are competed by a competitors that show homology to the consensus.

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 5) 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 14 matches (1 at 100% and
13 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 (Yang et
al, 1985) and the chicken βγ-globin enhancer (Muller and
Mehta, 1988) 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.
Furthermore, a 245 base pair fragment from upstream of the
HSV-1 IE gene was observed to have two 90% and one 100% topo
II consensus match; in vitro cleavage reactions showed that
these were cleaved by topo II, and when cleavage sites were
mapped in vivo, this region was found to be cleaved (Ebert,
S., Shtrom, S., and Muller, M., submitted). When examined
Table 5. Frequency of topoisomerase II consensus in enhancers.

<table>
<thead>
<tr>
<th>Enhancer</th>
<th>Real Sequence*</th>
<th>Random Sequence*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100%</td>
<td>90%</td>
</tr>
<tr>
<td>Adenovirus Ela (126 bp) (45)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>RSV LTR (296 bp) (Genbank)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Human Immunoglobulin (319 bp) (Genbank)</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Mouse Immunoglobulin (kappa) (479 bp) (43)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SV40 (strain 776) (183 bp) (39)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Chicken β-globin 3' enhancer (484 bp) (38)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>HSV-I IE gene 3 (230 bp) (40)</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>**Human β-globin 3' (700 bp) (Genbank, 41)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Matches of consensus to the actual sequence of enhancer.

*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.
further, this same region was found to be a powerful enhancer element.

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 (see above) 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 βA-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.
DISCUSSION

In order to derive a consensus sequence for DNA cleavage by an enzyme as promiscuous as topo II appeared to be, it was necessary to analyze a large number of sites, and from substrates with different base proportions. Failure to have done this would have resulted in a consensus which accurately described a particular data set, but had no universality for prediction of sites. These were the problems with the derivation of a consensus sequence for Drosophila topo II by Sander and Hsieh (1985); their result was true for their data set (16 sites of 70% A+T DNA), but could not predict the interaction of the enzyme with other DNA sequences (see Lee et al, 1989a,b; Spitzner and Muller, 1988). For this reason, 71 sites were catalogued from a range of substrates, as shown in Table 2. The drug m-AMSA was included in the reactions because drugs stimulate in vitro cleavages and appear essential for mapping sites in vivo (Yang et al, 1985; Rowe et al, 1986; Udvardy et al, 1986; Muller and Mehta, 1988). While cleavage patterns were not identical with and without drugs, they were shown to be similar (Udvardy et al, 1985), and here, 75% of the sites
were conserved with or without m-AMSA (Spitzner and Muller, 1988).

Although topo II is a homodimer, cleavage sites did not possess an axis of dyad symmetry. Subsequently, the major obstacle to the derivation of a consensus sequence was orientation of the strands to maximize the homology between the sites (see Table 3). This dilemma was addressed with a "by the bootstraps" operation, using a computer-assisted iterative method to derive a transitory consensus sequence, realign sites to match it, then derive a new temporary consensus and repeat the process. Eventually, there was a convergence towards certain bases and positions as consensus elements and it was necessary to select rules for their inclusion in a consensus sequence. Since the ultimate goal was a predictive model, it was decided that bases would be included only if they contributed to discrimination between sites and random sequences. The elements that were found to give the maximal prediction of sites versus nonsites are shown as the consensus sequence in Table 4; with it are shown the base proportions, by position, from the 71 sequences as best aligned to this consensus. As the graphs in Figure 11 illustrate, base proportions were essentially random at nonconsensus positions, and deviated from this only at the consensus positions.

A measure of the nonrandomness, or information at a position (reviewed by Storom, 1988) is its redundancy. The
percentage redundancy was determined for each position relative to the cleavage site and the results are shown graphically in Figure 12. This graph shows that all positions included in the consensus sequence have nucleotide proportions that deviate from background frequencies; it also indicates nonrandom occurrences at positions +10 and +11. However, inclusion in a consensus sequence of the pyrimidines that predominated at these two positions actually lowered the separation between sites and random sequences, so these were excluded. The consensus sequence derived covers 18 bases (10 bases 5' and 8 bases 3' of the topo II cleavage site); this fits the cleavage data observed (not shown), as restriction digestion of a substrate eliminates or dramatically lowers the cleavage of a site if the restriction site is within the consensus region, but does not affect topo II cleavage if the cut is outside the 18 base region. Within the 18 bases, 10 positions are specified; of these, 4 are single bases and 6 are degenerate. No positions were conserved in all sites, but the sites showed an average 70% homology to the consensus sequence. The degeneracy of the consensus sequence is reflective of the properties of the enzyme; cleavage sites have been observed at an average of one site per 25 bases (strong and weak sites), and the statistical probability of a 70% match to the consensus sequence is one per 25.9 bases (Spitzner and Muller, 1988). The promiscuity of topo II is
appropriate for its cellular function of chromosome segregation at mitosis (see above references), so that sequence constraints are not placed upon the DNA just to allow the decatenation.

Additional analyses suggested that the homology of sites to the consensus sequence is the major determinant for topo II cleavage. The di- and trinucleotide frequencies in the data base were only as expected by the mononucleotide proportions, and Markov chain analysis implied that the base occurrences at each position were independent of one another. An analysis of palindromes in sites showed no more occurrences than in random sequences, and Fourier analysis showed no significant periodicities of sequence elements. It was concluded that topo II sites were determined primarily by DNA sequence, and that if any other factors were involved it would require a much larger data set to show statistical significance of these factors. Further evidence that the recognition elements are at the cleavage site is the observation (data not shown) that a 32 base pair sequence containing a strong cleavage site can be cloned into a new location in a plasmid and still retain the same cleavage site. Additionally, cleavage sites have been identified about 10 bases from either end of DNA fragments (data not shown). Finally, Lee et al (1989b) demonstrated that topo II protected only a region overlapping the cleavage site from DNase I digestion.
Having derived a consensus sequence for topo II as a descriptive model for a particular data set of strong cleavages, it was of interest to determine whether this consensus model also had predictive value. Analysis of individual topo II sites suggested that these sites also exhibit homology to the consensus sequence (see Figure 13, for example); in addition it appeared that weak sites were simply sites with somewhat less homology to the consensus than strong sites. Further analysis suggested that a better model for topo II sites involved analysis of both strands of a cleavage site. A symmetric consensus sequence was too degenerate to be of great value for site description (see Figure 14), but comparison of both strands of a site independently to the asymmetric consensus sequence showed a good correlation between the total of the matches on both strands (The combined score or combined match) and DNA cleavage: strong sites had homology to the consensus on both strands, weak sites had good homology on one strand but minimal homology on the other, and sequences with a strand possessing little homology to the consensus were generally not cleaved even if the other strand showed homology. Thus, the topo II consensus sequence demonstrated both descriptive and predictive validity. Analysis of DNA for topo II sites by both cleavage reactions and by computer searches revealed a nonrandom distribution of sites. Topo II sites were relatively depleted in gene coding regions and enriched in
enhancer elements. The relevance of topo II sites to cellular functions of the enzyme has yet to be determined, however, the data suggest that topo II interacts with DNA in vivo at nucleosome-free sites, probably in or around nuclear matrix attachment regions (see references above), that are homologous to the consensus sequence.
Chapter IV

Methods for the quantitative analysis of topoisomerase II sites.

INTRODUCTION

Predicting sites with consensus sequences

In order to best understand the interaction between topo II and DNA it will be necessary to derive an optimal model for the description (and therefore, the prediction) of its recognition sites. The consensus sequence derived for DNA cleavage by Drosophila topo II (Sander and Hsieh, 1985) has validity in that sites with a very high degree of homology are indeed cleaved by the enzyme, but many sites with less extensive homology are also cleaved, while other sites of equal homology to these sites are not cleaved (Lee et al, 1989a,b). As demonstrated in the previous chapter, a more degenerate consensus sequence was derived for chicken topo II, using a larger, more diverse sample of sites. This consensus sequence was shown to be a good predictor of sites over a range of cleavage efficiencies. However, it is not an ideal model and suffers particularly in its quantitative...
value; sites of equivalent consensus sequence homology may exhibit very different strengths of cleavage (see above and Spitzner and Muller, 1988; Muller et al, 1988).

The problem of describing sites accurately with a consensus sequence is certainly not unique to topo II. Most DNA binding proteins exhibit a specificity of site selection, but few recognize a single unambiguous consensus sequence the way a restriction enzyme does. Instead, most of the DNA binding proteins studied (particularly transcription factors) display a range of activities from strong to weak binding (and indeed, this is a necessary aspect to their cellular functions). For example, the yeast TATA-binding protein TFIID was shown to bind promoters with and without the characteristic "TATA" consensus sequence (Hahn et al, 1989). In addition, a consensus sequence was derived for the binding of the transcription factor Sp1, but not all sequences bound by Sp1 are homologous to the consensus, nor are all consensus matches actually recognized by the protein (Gidoni et al, 1984; Dynan et al, 1986; Ares et al, 1987). A more extreme example is the yeast HAP1 activator, which was shown to bind sites with little apparent homology (Pfeifer et al, 1987). In fact, there are many cases of proteins that appear to recognize remarkably different DNA sequences (see also Baumruker et al, 1988; Costa et al, 1988; Johnson et al, 1987; and reviews by Stormo (1988) and Berg and von Hippel (1988); presumably, if
sufficient binding sites were examined a consensus sequence could be derived for these proteins. However, there are some serious problems associated with the use of consensus sequences; one is their arbitrary nature. For example, a position in which 52% of the sites had an adenine would be included in a consensus sequence, while a position with 47% of the sites having a cytosine might not, despite no statistical significance between those two proportions for a sample set. Perhaps more problematic is that one using a consensus sequence is assuming that all base matches (and mismatches) have an equivalent impact on the interaction between the protein and the DNA; in fact, the evidence suggests the contrary.

Clearly, not all mutations of sites have identical effects upon recognition, and single base changes generally result only in a partial loss of activity by a protein (see Stormo, 1988 and Berg and von Hippel, 1988 for reviews). Energetically, the major interactions between proteins and DNA in binding involve the formation of hydrogen bonds between the active site of a protein and the base pairs exposed in grooves of the DNA helix (Berg and von Hippel, 1988); however, additional factors, such as hydrophobic interactions, the steric fit and flexibility of DNA groove geometries, and counterion displacement also modulate binding affinities in sequence specific fashion. Therefore, not all base substitutions are equal, allowing for the
specification of binding sites with a wide range of affinity constants. This is of particular importance in cells, because it allows the use of a given protein at multiple sites, where the binding affinities of the sites regulates the activity of the proteins, and therefore also modulates the activities of cis-linked sequences. The best studied examples are of regulatory factors that bind sites in the promoters of genes to control the rate of transcription of the linked genes.

The first analysis of E. coli promoters suggested a conservation of nucleotides 10 bases 5' of the transcription start (the -10 region, Pribnow, 1975). The analysis of more promoters indicated that most matched the consensus sequence TANNNT, where N is any base (Hawley and McClure, 1983); however, this sequence alone can not specify a promoter. Identification of an additional region (at around -35) important for promoter activity, and the analysis of many E. coli promoters allowed the derivation and use of a consensus sequence that correlated well with the locations and the relative activities of the promoters (Hawley and McClure, 1983; Mulligan et al, 1984). Further advances were made with the advent of new computer technologies and the recruitment of computer scientists and mathematicians to tackle these biological problems. For example, complex algorithms were designed that utilized homologies to consensus sequences comprised of "words" that were
combinations of mono-, di-, and oligonucleotides; these models proved to be useful for the description and prediction of recognition elements (reviewed in Stormo, 1988). However, as mentioned above, there are limitations on the value of consensus sequences brought about by their composition of discontinuous events (matches are all or none), that do not permit the variability of affinities for binding sites that is demonstrated by proteins.

Predicting sites with matrix methods

Recently, methods have been developed for the analysis of binding sites which instead utilize the matrix of base proportions of a data set. In the consensus sequence method, if for some set of binding sites, at the -4 position 51% of the sequences had an adenine, 34% a guanine, 12% a thymine, and 3% a cytosine, an arbitrary decision would have to be made as to whether the consensus element at the -4 position should be called an adenine, a purine, or a "not cytosine". In a matrix method, the recognition element assigned to -4 would simply be the base proportions observed (or multiplied by some weighting measure, see below). Therefore, a binding site would be specified as a matrix of the base proportions of each position in the data set; the matrix associated with the chicken topo II binding sites
would be the 4 by 18 matrix from which the consensus sequence was derived.

The first use of a matrix method was by Stormo et al (1982) to identify *E.coli* translation initiation sites; they used a perceptron algorithm to derive a weight matrix which best discriminated between sites and nonsites. In general, once a matrix is derived for the positions of interest, a sequence is analyzed by matching each base of the test sequence to the score of that base in that position within the matrix. The scores for each position are summed to give a homology score for the whole test site; this is then compared to a threshold score above which signified a site and below which signified a nonsite in the data base used to derive the matrix. Matrix methods were utilized by Staden (1984) to make better predictions of *E.coli* promoter sites than could be achieved with consensus methods. Mulligan et al (1984) did a similar analysis using a matrix of probability values rather than simple base proportions and also got good results. An important advance in the analysis of sites was the calculation of the information content at a position (Schneider et al, 1986); this was essentially a measure of the deviation from random base occurrence at a position. This could be used both to decide what positions had above background information to include in an analysis and could also be used as a multiplier to weight base proportion matrices (see reviews by Stormo, 1988 and Berg
and von Hippel, 1988); matrix methods utilizing information content at positions were used to characterize cAMP receptor protein sites (Stormo and Hartzell, 1989; Berg and von Hippel, 1988).

In the statistical-mechanical model derived by Berg and von Hippel (see their 1988 review), the matrix model has been shown to be related directly to the actual binding energies of protein-DNA interactions. Given the assumption (usually correct) that bases contribute independently to the binding affinity of a site (and therefore, the base proportions observed for a data set represent the actual preferences of a protein), then the frequencies of individual base pairs will be proportional to the free energies of binding those bases, and the total free energy associated with binding the site will be the sum of the free energies of each base. Therefore, if true binding energies are known for sites, the matrix used to describe the sites will predict not only the locations of new sites, but also their binding constants. The limiting factor for the analysis of most recognition elements is acquiring a data base with a large number of sites and relative binding affinities. Therefore, topo II sites were analyzed using matrix methods and relative cleavage strengths. The results indicate that this model makes accurate quantitative predictions of topo II sites.
MATERIALS AND METHODS

Materials. Reagents were as described above.

Cleavage reactions. Topoisomerase II cleavage reactions were performed on the specified 5'-32P-end-labeled DNA fragments as described in Spitzner and Muller (1988) and Muller et al (1988). Briefly, reactions (final volume, 20 μl) contained 20,000 dpm of end labeled DNA (about 3 ng) in cleavage buffer (30 mM tris-HCl pH 8.0, 60 mM KCl, 8 mM MgCl2, 3 mM ATP, 5 mM 2-mercaptoethanol, and 40 μg/ml bovine serum albumin), and 10 units purified chicken topoisomerase II (Muller et al., 1988) plus 50 μg/ml 4'-(9-acridinylamino)methanesulfon-m-anisidide (m-AMSA) where specified. Reactions were incubated for 10 min at 30° C, then terminated with sodium dodecyl sulfate to 1%. Samples were treated with proteinase k, ethanol precipitated, resuspended in sequencing dye, boiled 3 min, then approximately 10,000 dpm were loaded onto sequencing gels along with 10,000 dpm of Maxam and Gilbert sequencing markers. Densitometric scans were performed with an LKB soft laser densitometer.
RESULTS

Representing topo II sites with base proportion matrices.

We analyzed topoisomerase II cleavage sites on a large number of diverse DNA sequences and derived the consensus sequence shown in Table 6 (Spitzner and Muller, 1988). Despite the dimeric nature of topoisomerase II, most cleavage sites did not possess an axis of dyad symmetry, and the consensus sequence is clearly asymmetric. However, it became clear that the frequency of topoisomerase II cleavage at a site was, in fact, related to the degree of homology to the consensus sequence on both strands of the site, and sequences with a good consensus match on one strand but with few positions matched on the opposite strand were cleaved poorly, or not at all (Spitzner and Muller, 1988). Additionally, strong topoisomerase II sites have more consensus sequence matches on both strands of a site than do weaker enzyme 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 using
Table 6. Nucleotide proportions in the topoisomerase II consensus sequence.

<table>
<thead>
<tr>
<th>Position</th>
<th>Nucleotide proportion</th>
<th>Consensus Base</th>
<th>Sequence Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>-10</td>
<td>0.423</td>
<td>0.141</td>
<td>0.268</td>
</tr>
<tr>
<td>-9</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-8</td>
<td>0.141</td>
<td>0.352</td>
<td>0.155</td>
</tr>
<tr>
<td>-7</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-6</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-5</td>
<td>0.141</td>
<td>0.521</td>
<td>0.141</td>
</tr>
<tr>
<td>-4</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-2</td>
<td>0.183</td>
<td>0.141</td>
<td>0.521</td>
</tr>
<tr>
<td>-1</td>
<td>0.056</td>
<td>0.366</td>
<td>0.085</td>
</tr>
<tr>
<td>Cut site</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+2</td>
<td>0.282</td>
<td>0.113</td>
<td>0.535</td>
</tr>
<tr>
<td>+3</td>
<td>0.085</td>
<td>0.113</td>
<td>0.451</td>
</tr>
<tr>
<td>+4</td>
<td>0.127</td>
<td>0.056</td>
<td>0.042</td>
</tr>
<tr>
<td>+5</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+6</td>
<td>0.155</td>
<td>0.282</td>
<td>0.113</td>
</tr>
<tr>
<td>+7</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+8</td>
<td>0.070</td>
<td>0.352</td>
<td>0.113</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>C</td>
<td>G</td>
</tr>
</tbody>
</table>
a matrix based on the observed frequencies of nucleotides at each position relative to the cleavage site. This approach was used by determining which $n$ positions had nonrandom base frequencies and then constructing a $4 \times n$ matrix of the proportions of each nucleotide at each position (see Table 6). 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. As a number of different methods for analyzing topo II sites using the consensus sequence information will be used and contrasted, the various methods are described below, with examples.

**Homology as determined by direct comparison to the consensus sequence**

Sites are analyzed by positioning the consensus sequence (see Table 6) over the cut site, as shown in Table 7. The enzyme cut site is between positions -1 and +1, with minus positions 5' and plus positions 3' of the cleavage site. The "combined match" (referred to previously as the "asymmetric consensus sequence" method in Spitzner and Muller, 1988) sums the number of consensus sequence matches on one strand of a potential cleavage site with the number of matches to the same consensus sequence on the opposite strand of the site. The consensus sequence is 18 bases in length, but matches are scored only to the 10 non-"N"
Table 7. Homology calculations at a sample site.

<table>
<thead>
<tr>
<th>Consensus Sequence</th>
<th>r n Y n n C n n g Y n g t n Y n y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotide proportion (Table 1)</td>
<td>169 352 521 183 493 07 113 127 282 113</td>
</tr>
<tr>
<td>Consensus Sequence Position</td>
<td>-10 -8 -5 -2 -1 2 3 4 6 8</td>
</tr>
<tr>
<td>Test Sequence</td>
<td>Top 5' T C G T C G A A T A T C A A C T G G C A T T G 3'</td>
</tr>
<tr>
<td></td>
<td>Bottom 3' A G G C A G C T T A T A G T I T G A C C G T A A C 5'</td>
</tr>
<tr>
<td></td>
<td>352 451 775 085.535 493 521 521 141 141</td>
</tr>
<tr>
<td>Y n Y n T G n Y G n n C n n n r</td>
<td></td>
</tr>
</tbody>
</table>
consensus positions. In the combined match method, bases which match the consensus sequence are given the same weight whether the position matched is a single base (like the +4 T) or two bases (like the -1 pyrimidine). After the number of consensus sequence matches to the (arbitrarily chosen) top strand is determined (4 for the site in Table 7, where matches are shown as upper case letters and mismatches as lower case letters), the consensus sequence is aligned across the corresponding cleavage site on the bottom strand, which is 4 bases 5' of the top strand site since topoisomerase II makes a double strand break in DNA leaving a 4 base 5' overhang. The bottom strand is scored for consensus sequence matches as above; the sequence in Table 7 scores 7 of 10 matches. The combined match is the sum of the number of consensus sequence matches on both strands of a site; for the sequence in Table 7 there are 4 matches on the top strand and 7 matches on the bottom strand, which equals a combined match of 11.

Homology as determined by using the consensus sequence matrix

The information within the topoisomerase II consensus sequence can also be represented as a 4 by 18 matrix of 0's and 1's, as shown in Table 6. The rows are the sequence positions relative to the cleavage site and the columns correspond to the bases A, C, G, and T, respectively.
Matches to the consensus sequence (such as the -2 G and the +3 G or T) are scored as 1. Non-consensus sequence positions within the matrix are 0's, as are mismatched positions such as -10 C or T and +4 A, C, or G. For the combined match method, all consensus positions are weighted evenly, but algorithms utilizing the match of a potential topoisomerase II site to a matrix are easily adapted by changing the values in the matrix. It is convenient to analyze consensus sequence matches to all potential cleavage sites on a known DNA sequence with the aid of a computer program (we used the EDEN Genesys program). At each site, the first base (10 bases 5' of the cleavage site) is compared to the first (-10) matrix position and this position is given the value that the base has in the matrix (1 if it is A or G, 0 if the base is C or T). Next, the -10 position on the other strand is evaluated by examining the position which is 23 bases 3' of the top strand -10 position and taking the complement of the top strand base to yield the base that exists in the bottom strand -10 position. This is scored as above and the value, either 0 or 1, is added to the score for the top strand -10 position. Next, the -9 positions on each strand are examined, but these will always score 0's because this is not a consensus sequence position. Then, the -8 position is examined for each strand and these scores are added to the running total. After all 18 consensus positions (10 of them non-"N") are
evaluated for both strands, the running total is the combined match.

**Homology as determined by the matrix mean**

This method is similar to the method just described, except that the matrix contains instead the actual base proportions from the previously derived topoisomerase II consensus sequence (Spitzner and Muller, 1988), shown in Table 6. Again, positions are evaluated one by one and each receives the value in the matrix corresponding to the base at that position. In analyzing the example site in Table 7, first the -10 top strand T is scored as 0.169 (from the matrix) and the bottom strand C is scored as 0.141. The -9 position is not a consensus position, so it adds nothing to the score. Next, the -8 positions on each strand are scored. The top strand C gets a 0.352 and the bottom strand A gets 0.141; these values are added to the running total. Once all 18 positions have been scored for each strand, the total of the 10 top strand and 10 bottom strand non-"N" scores is divided by 20 to yield the average base proportion per position. We refer to this value as the "matrix mean." The matrix mean for the site in Table 7 is 0.322; the highest possible score is 0.499.
Homology as determined by the unique score

The "unique score" is the number of the 10 non-"N" consensus positions which is matched on one strand or the other at a site. The unique score is calculated by summing the number of positions matched for each of the strands (the combined match) and then subtracting the number of duplicates (positions matched on both strands). For example, the sequence in Table 7 matches the -8, -5, -1, and +6 consensus positions on the top strand and the -5, -2, -1, +2, +4, +6, and +8 positions on the bottom strand. The positions matched on at least one strand are then -8, -5, -2, -1, (+)2, 4, 6, and 8, so the unique score is 8 out of 10. The match to the +6 position on both strands does not increase the unique score in comparison to a sequence where only one match was found to this consensus position; only matches to the -10 and +3 positions would increase the unique score. The unique score may be calculated using the consensus sequence matrix (0's and 1's) and an algorithm similar to the one above. At each of the 10 non-"N" consensus positions, call the top strand match (0 or 1) x and the bottom strand match y. The maximum value of (x,y) is determined at each position. This maximum will have a value of 1 if either the top strand, the bottom strand, or both possessed a match to the consensus sequence matrix at that position, and a value of 0 if neither strand was a match (that is, if the matrix value on each strand was 0 for
the position analyzed). The maximum values for each of the 10 consensus positions are summed to yield the unique score (which can range from 0 to 10). In summary, a number of ways to evaluate the homology of a site to the topoisomerase II consensus sequence have been presented; these represent three alternatives for scoring a site. The combined match is the total number of consensus sequence matches on both strands of a site, the matrix mean is the average match of a site to the nucleotide proportions used to generate the consensus sequence, and the unique score is the number of consensus sequence matches on one strand or the other at a site.

**Analysis of topo II sites using matrix methods**

In the first analysis comparing accuracy of site prediction for consensus sequence versus matrix methods, a slightly different matrix mean method was used: instead of using the base proportion at a position directly, its natural logarithm was substituted. This has the result that final matrix score (after computing the antilogarithm of the total) is equivalent to multiplying rather than adding successive base proportions; since base frequencies are essentially probabilities of base occurrence, and probabilities of related events are computed by multiplying probabilities of each individual event, the result of this analysis procedure is basically the probability that the
test sequence is a topo II site as compared to the optimum consensus site. In the analysis (Table 8), cleavage sites were analyzed on 1200 bases of DNA and a total of 66 cleavage sites were observed; of these, 14 were strong sites. The 1200 bases of sequence were subject to analysis by a consensus sequence method (combined match, using a stringent score of 11 as the threshold) and by the probability calculation using a threshold antilog value of 3 x 10^{-6}; analogous symmetric consensus sequence methods (see Chapter III) were also examined. It was clear that all methods were accurate for detection of strong sites (13/14), and all methods predicted roughly half of the weak sites; the differences were that, for either method, the asymmetric consensus sequence was more accurate in terms of predicting sites and nonsites. There was no major advantage between the matrix and the consensus method: one predicted more sites accurately, the other had fewer false positives. The thresholds selected affect the result quite a bit; if the cutoff for the asymmetric probability method is lowered to 1 x 10^{-6} then 39 of the 66 sites are predicted, but the number of false positives rises from 15 to 39. If the cutoff is raised to 10 x 10^{-6}, only 16 of the 66 sites are predicted, but the number of false positives drops to 3. In general, thresholds are chosen to try to maximize the number of sites predicted and minimize the number of false positives.
Table 8. Comparison of the symmetric and asymmetric consensus sequences to the data base.

<table>
<thead>
<tr>
<th>Consensus Sequence:</th>
<th>Total Number of Sites Analyzed</th>
<th>Actual Number of Topo II Sites:</th>
<th>Number of Sites Accurately Predicted</th>
<th>Number of Sites Falsey Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comparison Based on % Match to Consensus Sequence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Comparison Based on Proximity Calculations</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Asymmetric Consensus Sequence: 5' RNYNNCNNYG^* N G 6', TNYNY 3'  
Symmetric Consensus Sequence: 5' RNNNNYRNRY^* V B 6', RYNYRNNNY 3'  
(V = not T; B = not A)

This value corresponds to the total number of nucleotides in the data base used in this comparison.

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 results in Figure 15 compare the predictive values of the same methods as above on the cleavage sites detected on a kDNA fragment. Here, the asymmetric matrix method is the superior predictor: 12 of 12 sites with no false positives. As more sequences were examined, it was found that simply adding the matrix values directly (rather than their logarithms) discriminated better between sites and nonsites, so the matrix mean is used instead of the probability calculation.

**Predicting the strength of topoisomerase II cleavages**

The next question addressed was whether the strength of cleavage (or relative frequency of cleavage at a site) could be predicted using the consensus sequence information. For example, when the reactivity of topoisomerase II to alternating purine-pyrimidine sequences was examined, the enzyme was found to make stronger cleavage bands on (GT)_n than on (GC)_n repeats (Spitzner, Chung, and Muller, 1990). The results of analyzing both sites by the different methods are presented in Table 9. The "best strand match" (footnote c) did not distinguish between the sites. The combined match was lower for the stronger cleavage site (12 versus 14 for the (GC)_n site). The unique score was greater for the stronger site (8) than the weaker site (7); they differ by nearly 1 standard deviation in score (see Table 8). The
### A

<table>
<thead>
<tr>
<th>Consensus Sequence</th>
<th>Total No. of sites/100^\text{*}</th>
<th>No. of Sites Accurately Predicted</th>
<th>No. of Sites Falsely Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis 1: Based upon % match to Consensus Sequence</td>
<td>Asymmetric (see table 2)</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>Symmetric (see table 3)</td>
<td>12</td>
<td>9</td>
<td>2^*</td>
</tr>
<tr>
<td>Analysis 2: Based upon probability calculations (see results)</td>
<td>Asymmetric (see table 2)</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Symmetric (see table 3)</td>
<td>12</td>
<td>10^*</td>
<td>0</td>
</tr>
</tbody>
</table>

^*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).

^A total of twelve cleavages were selected for this analysis and are marked below in Fig. B.

^The sites which we failed to predict were relatively weak sites; for example in this case, sites 6 and 9 (Fig. B) were not predicted according to our threshold conditions (see text).

^These two sites were predicted by a favorable match to consensus, but were not detected in the experiment in Fig. B.

### B

![Figure 15](image)

**Figure 15.** Comparison of symmetric and asymmetric topo II consensus sequences to actual sites. Panel A. Site predictions. Panel B. Sequenced topo II sites on kDNA fragment (sites are numbered 1 to 12).
Table 9. Unique score is the better measure of topoisomerase II cleavage strength than the combined score

Vertebrate topoisomerase II consensus showing critical nucleotides

<table>
<thead>
<tr>
<th>-10</th>
<th>-8</th>
<th>-5</th>
<th>-2</th>
<th>-1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>C</td>
<td>N</td>
<td>N</td>
<td>G</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>G</td>
<td>Y</td>
<td>N</td>
<td>G</td>
<td>T</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>T</td>
<td>K</td>
<td>G</td>
<td>N</td>
<td>Y</td>
<td>G</td>
<td>N</td>
</tr>
<tr>
<td>G</td>
<td>N</td>
<td>C</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>R</td>
<td></td>
</tr>
</tbody>
</table>

8 6 4 3 2 | -1 | -2 | -5 | -8 | -10

(a) Match of (GC)₈ to the topoisomerase II consensus sequence

Best strand = 7/10⁰; combined match = 14/20⁹
Unique score = 7/10⁰
Matrix mean = 0.309⁰

(b) Match of (GT)₈ to the topoisomerase II consensus sequence

Best strand = 7/10⁰; combined match = 12/20⁹
Unique score = 8/10⁰
Matrix mean = 0.318⁰

*The site of topoisomerase II cleavage is marked at the arrows. 'N' indicates that no nucleotide is preferred at the position. R is purine, Y is pyrimidine and K is G or T. The consensus sequence is from Spitzner and Muller (1988).

b(GC)₈ is cleaved weakly by topoisomerase II, whereas (GT)₈ is cleaved much more frequently (Spitzner et al., 1989).

'best strand match' is the conventional application of the consensus sequence; it is the number of consensus positions matched by whichever DNA strand at a given site has greater homology to the consensus sequence.

The combined match is the sum of the top strand matches (to 'non-N' positions of the consensus sequence) plus the bottom strand matches, which is 7 + 7 = 14 for (GC)₈ and 7 + 5 = 12 for the (GT)₈ DNA sequence.

The unique score is the number of consensus positions which are matched on at least one strand. For example, the unique score for the (GC)₈ sequence is 7/10 because 7 of the 10 consensus positions are matched (black dots above the sequence mark the matched positions) while 3 are not matched (−8, 2 and 4).

'The Matrix Mean' corresponds to the sum of nucleotide proportions of all non-N positions in the data base (see Table 1) divided by 20.
matrix mean was slightly larger for the stronger topoisomerase II site, but by only a small fraction of 1 standard deviation.

An illustration of the scoring model (defined as a matrix mean of at least 0.28 where the unique score then determines strength of cleavage) is shown in Figure 16. A 121 base pair sequence from the HSV-1 immediate early gene 3 promoter/enhancer region was reacted with topoisomerase II in the absence or presence of m-AMSA. The drug did not alter the specificity of the enzyme, but it enhanced the intensity of the topoisomerase II cleavage bands. Due to the band intensities in the presence of m-AMSA, the autoradiograph was exposed for a rather short time. This is one reason why band intensities are measured as "relative cleavage strength" rather than absolute, because these sites (a and b) are actually fairly strong sites given the absence of drugs (cleavage strengths of 4 and 2, respectively). Cleavage bands are often better observed by densitometric scans of the gels because cleavage band intensities do not always reproduce accurately as photographs of gels; the scan for this gel (of the lane lacking m-AMSA) is shown below the gel in Figure 16. The bands are clearly quite intense compared to the background of the autoradiograph. The strongest band, site (a), scored 10 out of 10 unique matches and had a combined match of 13, while the moderately strong site (b) had a unique
Figure 16. Topoisomerase II cleavage on an HSV-1 DNA fragment is predicted by the unique score model. A 121 base pair 5' end-labeled AvaI-NcoI fragment (labeled at the AvaI site) from the promoter/enhancer region of HSV-1 immediate early gene 3 was reacted with topoisomerase II in the absence or presence of m-AMSA (indicated by the lanes), as described above. Reaction products were analyzed on a 12% polyacrylamide sequencing gel (with a Maxam and Gilbert G ladder). The DNA sequence is shown below the sequencing ladder; below this is the unique score for every second potential cleavage site. All sites scoring above both the unique score threshold of 6 and the matrix mean threshold of 0.28 are indicated by a box drawn around the unique score; these sites are the two cleavage sites (a and b) and two sites that were not cleaved by topoisomerase II in the absence of m-AMSA (and these false positives have low unique scores). Due to differences in mobility between topoisomerase II cleavage products (with a 3' hydroxyl) and chemical sequencing markers (3' phosphates), the actual cleavage site is one base 5' from where the cleavage site appears to be in the DNA sequence. A densitometric scan of the cleavages without drug is shown. The sites labeled (a) and (b) are marked and above each site is the unique score and, bracketed, the combined score.
score of 9 and a combined match of 12. In the absence of m-AMSA, no other cleavage bands were seen on this fragment and this is essentially as predicted by the model; background unique scores are shown below the sequencing gel for potential cleavage at every second base along the sequence. Nearly every potential site above threshold for the unique score failed to meet the other criterion of a matrix mean greater than or equal to 0.28; the sites which met both thresholds are indicated by a box around the unique score. These boxed values are the two cleavage sites (a and b) and two false positives that were predicted to be weak sites (and one of these was weakly cleaved when m-AMSA was included). In general, most false positives are predicted as weak sites (unique score of 6 or 7). Many of these may have the potential to be real cleavage sites because cleavage at weak sites is often suppressed when sites are flanked by strong sites (which presumably compete better for the enzyme; if strong sites are eliminated from a DNA fragment then weaker appear or increase in intensity, Ebert, Spitzner, and Muller, unpublished results).

The results of additional experiments are shown in Figure 17. Figure 17A shows densitometric tracings of topoisomerase II cleavage on a pUC12 lacZ gene sequence both in the absence and presence of m-AMSA; the results are qualitatively the same (although the gain was adjusted higher on the densitometer for the cleavage bands without
Figure 17. The unique score model makes accurate predictions of topoisomerase II cleavage sites and frequencies in the absence or presence of m-AMSA. A, B, and C are all densitometric scans of sequencing gels containing topoisomerase II cleavage reactions on the DNA described. In each case, all of the detectable peaks, representing the cleavage bands, are labeled with letters, and above each is the unique score for the site. Above the unique scores in A are the combined matches in brackets. Panel A. Two strong sites were identified in the lacZ gene of pUC12 (see Muller et al, 1988), and scans are shown for the sites in both the absence and presence of m-AMSA, as marked. The two strong sites are labeled (a) and (b). Panel B. A uniquely end labeled fragment (89 base pair NcoI-SphI) from the 5' region of immediate early gene 3 of HSV-1) in the presence of m-AMSA; sites are labeled a-h on the scan. Panel C. Cleavages were performed with m-AMSA on a kinetoplast DNA sequence (5' end labeled Xho1-Hinfl, 723 base pairs, labeled at the Xho1 site) prepared from C. fasciculata minicircles. Sites are labeled a-m; site (i), labeled with an asterisk, was also observed in the absence of topoisomerase II, so this is an artifact, not a cleavage band.
drug); the strongest site (b) had a unique match of 9 and the somewhat weaker site scored 8 unique matches. The combined match was not as accurate a predictor, as site (a) scored higher than the more frequently cleaved site (b).

Figures 17B and C demonstrate the validity of the unique score model even when there are many clustered cleavage sites on a fragment. Figure 17B shows m-AMSA enhanced topoisomerase II cleavages on a kinetoplast DNA sequence. The strongest cleavage band is site (m) which has a unique score of 10 out of 10. The next strongest sites, (b) and (c), score 9 of 10. Site (i) is an artifact because the band appeared in the absence of enzyme; thus it is not a cleavage band. Site (a) scores 8 and is cleaved moderately strong, while the sites with a unique score of 7 show weaker cleavages. Note that sites (e) and (j) are cleaved less frequently than would have been predicted purely by unique score; it is important to score all sites on a fragment first and assign relative cleavage strength values to the sites with the highest unique scores and to unclustered sites before examining sites which are proximal (up to 10 bases away) or flanked by strong sites. Sites such as (e) and (j) are consistently cleaved less frequently than the score predicts because both are near to strong sites and have strong sites located both 3' and 5' of them. Similar results are found for cleavage of an HSV-1 enhancer sequence, as shown in Figure 17C. The strongest
topoisomerase II sites, (b) and (g), have unique scores of 9 (the highest on the fragment). The next strongest sites are (c) and (h) with 8 unique matches (note, neither site is closely surrounded by other, stronger sites). In contrast, site (f) is in a cluster of sites and site (a) is located only a few bases from site (b) and is very near the 5' end of the fragment. Thus the inclusion of m-AMSA in topoisomerase II cleavage reactions serves mostly to enhance the strength of cleavage sites and does not change the rules for enzyme recognition of sequence in any major way. An independent data base of 60 cleavage sites identified in the presence of the drug showed that the weak sites had an average unique score of 7.2, while the strong sites averaged 8.7.

Analysis of sites without drugs

It was important to show that the unique score model using the consensus sequence derived from sites in the presence of m-AMSA was also valid for sites with no drugs added. It was possible that m-AMSA greatly altered the sites recognized by the enzyme; however, this was not likely since 75% of all sites observed in the presence of m-AMSA were also sites in its absence. In another cleavage experiment, in the absence of any drugs, topoisomerase II was reacted with a DNA fragment from the upstream region of
an archaebacterial gene (see Spitzner and Muller, 1988) in the absence of any drugs. The densitometric scan is shown (Figure 18). Site (c) is the strongest cleavage site we have observed by topoisomerase II alone, and its sequence is almost perfectly symmetric with respect to consensus sequence matches (10 on one strand and 7 on the other; the best possible is 10 on one strand and 8 on the other because the +2 and +3 positions cannot be matched on both strands). Site (c) is also predicted to be the strongest site on the fragment by its unique score of 10 out of 10. Site (b) also scores above threshold by both models, but matches only 3 positions on one of its strands, so despite its matrix mean of 0.314 and a unique score of 10 it was cleaved only very weakly (as observed in Spitzner and Muller, 1988). Site (a) has a unique score of 8 and is moderately cleaved as the model predicts; sites (d), (e), and (f) have the unique scores indicated and show roughly the moderate to weak cleavages expected. Clustered sites such as these generally show lower frequencies of cleavage than anticipated purely from the equations for the models. Again, all cleaved sites scored above threshold for the unique model (a score of 6); in fact, all sites observed on this fragment had unique scores of at least 7 (and in general, sites which have unique scores less than 7 are cleaved weakly if at all). Examination of uncleaved sites showed that the majority were indeed predicted to be nonsites by failing to meet the
Figure 18. Topoisomerase II cleavages on an archaebacterial sequence. The DNA fragment is described in Spitzner and Muller (1988). A cleavage reaction was performed in the absence of any drugs as described above. A densitometric scan of the sequencing gel is shown, with cleavage bands labeled a-f; above each topoisomerase II site the unique score is shown, along with the combined score in brackets. Site (c) is the strongest site we have observed with enzyme alone and its sequence is highly symmetric with respect to consensus sequence matches:

```
5' A C T T G C G G G T | A G T T A C C C A T G A T T
3' T G A A C G C C C A T C A A | T G G G T A C T A A
```

The cleavage occurred between the tenth and eleventh bases (|) reading 5' to 3' on each strand. The background (uncleaved site) unique scores are mostly 5, 6, or 7, where the sites scoring 6 or 7 in most cases (>80%) are predicted not to be sites by the other rules given (in the text) that govern topoisomerase II cleavage.
cleavage criteria of both the matrix mean and unique score thresholds (these sites are not shown), and of the few false positives, most were located within a few bases of cleavage sites with higher unique scores.

Thus the unique score appears to be a valid model for quantitative predictions about topoisomerase II sites, and predictions about strong sites are made with the highest degree of accuracy; the few exceptions observed generally concern weak sites and lower threshold scores.

Compilation of a data base with 65 no drug topo II sites

To investigate the generality of the findings for sites with enzyme alone, a new, independent sample of topo II sites was derived from reactions lacking any drugs. Topoisomerase II cleavage reactions were performed on a variety of DNA sequences from mammalian genes and promoters, bacterial plasmids, trypanosome kinetoplasts, and mammalian viral genes and promoters without the use of any drugs such as m-AMSA or VM26 to enhance cleavages, and a total of 65 sites were sequenced (see Spitzner and Muller, 1989); these sites represented a wide range in cleavage strengths. For each site, prior to any analysis of sequence, the relative cleavage strength was recorded. Given our current methodology, cleavage strength can not be scaled in an absolute manner; rather, the measure of cleavage strength
depends upon the amount of radioactivity loaded onto a sequencing gel, the exposure time of the X-ray film, and even the conditions for development of the film. In addition, most cleavage bands contain only a small fraction of the total counts loaded onto a gel, especially in the absence of drugs that enhance cleavages. Within a sequence, however, the relative cleavage strengths are reproducible, and from experiment to experiment the scoring of relative cleavage strengths remains consistent. In general, DNA fragments averaged about 300 base pairs in length, and 10,000 counts per minute of both the cleavage reaction and the sequencing markers were loaded onto a sequencing gel. Extremely intense cleavage bands, those more intense than the chemical sequencing markers, are scored as "5's". Cleavages which are on the order of the intensities of the sequencing markers are scored as "4's". Bands which are barely detectable above background are "1's", and "2's" are somewhat stronger. Moderately strong sites are scored as "3's"; sites scoring 3 or more are considered strong sites, such as were used to derive the consensus sequence, while 1's and 2's are weak sites. The 65 site sample contains 42 weak sites and 23 strong sites. Analysis of the sequences examined indicate that there is, on average, one topoisomerase II site per 22 bases of DNA sequence (from weak to strong), one strong site (scores >3) per 62 bases, and one very strong site (a "5") only once per 700 bases.
Comparison of the independent sample of 65 topoisomerase II sites with the original 71 sites used to derive the consensus sequence.

The topoisomerase II consensus sequence data was originally derived from the 71 cleavage sites identified in the presence of m-AMSA, so the 71 sites do, of course, show homology to the consensus sequence. An important question to address was whether this consensus sequence also predicted cleavage sites in an independent sample of enzyme sites that were identified from reactions which did not include any topoisomerase inhibitors. There were several possible reasons why a new sample of data might not show the same match to the consensus sequence as the original set of data. If the consensus sequence contained information gained from capitalizing on chance associations within the 71 site sample or if weaker cleavage sites found within the new sample were not homologous to the 71 strong sites then the original consensus sequence would not be expected to apply equally well to the independently collected 65 topoisomerase II sites. Additionally, the set of cleavage sites identified on a sequence in the presence of m-AMSA is not always identical to the set identified in the absence of drugs. However, if the new data base of sequences did show significant homology to the original consensus sequence, this would be verification that the consensus sequence can be used to predict vertebrate topoisomerase II sites in the
absence or presence of drugs for both strong and weak sites.

Many different methods for characterizing the consensus sequence data from Spitzner and Muller (1988) were examined and the most useful methods are described above. These methods were used to compare the scores of the 71 sites obtained in the presence of m-AMSA with the scores of the 65 sites not enhanced by inclusion of drugs. The various methods described utilize information from both strands of a site, avoiding the potential problems associated with the need to align the sequences before analysis, as was necessary to derive the asymmetric topoisomerase II consensus sequence (Spitzner and Muller, 1988). Comparison of scores for the two samples and for nonsites, using the same consensus sequence information, is shown in Table 10. These results suggest that it is quite reasonable to use the original consensus sequence to analyze the new data sample; thus the consensus sequence is applicable to enzyme sites identified in the absence of any drugs.

**Discrimination between topoisomerase II cleavage sites and nonsites**

The 65 topoisomerase II sites obtained in the absence of drugs contain 560 nonsites, defined as sites flanking the cleavage sites which were not cleaved by the enzyme in the absence of drugs. (When these nonsites were compared with
Table 10. Comparison of data bases.

<table>
<thead>
<tr>
<th></th>
<th>71 sites a (+ drugs)</th>
<th>65 sites b (- drugs)</th>
<th>23 strong sites c (out of 65)</th>
<th>nonsites d</th>
</tr>
</thead>
<tbody>
<tr>
<td>combined match</td>
<td>11.3 (1.96)</td>
<td>10.2 (2.46)</td>
<td>11.2 (2.17)</td>
<td>7.6 (2.07)</td>
</tr>
<tr>
<td>average (SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>matrix mean</td>
<td>0.315 (0.036)</td>
<td>0.294 (0.047)</td>
<td>0.314 (0.041)</td>
<td>0.246 (0.039)</td>
</tr>
<tr>
<td>average (SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unique score</td>
<td>8.1 (1.2)</td>
<td>7.6 (1.6)</td>
<td>8.2 (1.3)</td>
<td>6.0 (1.5)</td>
</tr>
<tr>
<td>average (SD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a The 71 topoisomerase II cleavage sites, obtained in the presence of m-AMSA, were used to drive the consensus sequence (Spitzner and Muller, 1988).

b The 65 sites, obtained in the absence of drugs, are described in the Results; the 23 strong sites (cleavage strength greater than or equal to 3) are a subset of this data base.

c The nonsites are those sites in the two databases above which cleavage was not observed.
another set of nonsites which did not closely flank topoisomerase II sites the consensus match distributions were essentially identical). Many methods for characterizing the consensus match were explored, including procedures using one or both strands of sites, transforming matrix values to logarithms, using a variety of weighting systems, and testing alternative methods (arithmetic and geometric) for averaging values from the two strands. The object was to find a combination of computational method and threshold value that maximized the identification of cleaved sites while minimizing the prediction of cleavage at nonsites ("false positives"). Specifically, for each value of the threshold, the percentage of misclassifications was computed separately for sites (i.e., percentage of non-predicted sites) and nonsites (percentage of false positives). These two percentages were then averaged arithmetically to yield a discrimination score. The method that yielded the best discrimination value was the matrix mean (see above). Analysis of variance comparing the average matrix means for sites and nonsites indicated that the value was significantly higher for sites than for nonsites (F(1,623) = 79.29, p<.001). This relationship is depicted in Figure 19 (A and B): enzyme sites and nonsites show different, although somewhat overlapping distributions. Using the matrix mean method, a threshold of 0.28 yielded the best discrimination between sites and
Figure 19. Optimizing discrimination between sites and nonsites. Graphs A and B show the frequency distributions of the cleavage sites (X) and nonsites (+) with respect to
Figure 19 (continued)

matrix mean scores. Each point represents the number of sites scoring within a matrix mean range of 0.01. The optimization curve (open circles) plots the arithmetic average of the percentage of cleaved sites misclassified and the percentage on nonsites misclassified (with respect to the given matrix mean value as the threshold). For example, at the matrix mean threshold of 0.28, 45 of the 57 cleaved sites that had unique scores of at least 6 were classified correctly, so 12 of these 57 (21.1%) were misclassified. Likewise, 99 of the 351 nonsites with unique scores greater than or equal to 6 (28.2%) were misclassified; therefore, the average of the percentages of misclassifications is 24.6. Panel A shows the optimal criteria of only allowing nonsites and cleavage sites which have unique scores of at least 6. Panel B shows all sites. Panel C is a bar graph of the distribution of frequencies of all 65 cleavage sites (height of hatched bars) and of the 23 strong topoisomerase sites (shaded bars) within each of the 0.01 matrix mean intervals. The selected threshold value for the matrix mean (0.28) is indicated by a vertical line; 21 of 23 strong sites have scores equal to or exceeding this value.
nonsites (Figure 19B). Discrimination between sites and nonsites was improved by the additional requirement that a site must have a unique score (as defined above) of at least 6 (out of 10). Figure 19A depicts the behavior of the discrimination measure as the threshold is varied; in this panel, only sites meeting the second criterion (unique score greater or equal to 6) are included. At a matrix mean threshold of at least 0.28, 69% of all cleavage sites were classified correctly, as were 84% of all nonsites (a false positive percentage of 16%). Lower threshold values allowed too many false positives, while higher values did not detect as many true sites. For example, at a matrix mean threshold of 0.305, the false positive rate dropped to 6%, but only 52% of the 65 real sites were included. Figure 19C shows that the same criteria predicted strong sites better than weak sites. Using the threshold values of unique score of at least 6 and matrix mean of at least 0.28, 21 of the 23 strong sites in the 65 site sample (91%) were included. More stringent cutoffs still allowed good prediction of strong topoisomerase II sites. If the matrix value was raised to 0.295, 78% of strong sites were classified correctly and 90% of nonsites were classified correctly. At a matrix mean of 0.305, only 6% of nonsites were misclassified, while 65% of the strong enzyme sites were still predicted accurately. Therefore, the thresholds set depend upon whether one wants to maximize the number of
cleavage sites classified correctly or minimize the number of false positives predicted, and whether it is necessary to identify all topoisomerase II sites or only the strongest sites.

**Cleavage strength is correlated with consensus match**

To determine whether the consensus match of the sites was proportional to the strength of cleavage at the site, statistical analyses were performed using the cleavage strengths and matrix means or unique scores for all 65 topoisomerase II sites (obtained without drugs) and the sample of 560 nonsites. Analyses utilized the EDEN Genesys computer program (Joseph Spitzner, TEAM Associates, 487 Catawba Ct., Westerville, Ohio). Analysis of variance indicated that the average matrix mean varied among nonsites, weak sites, and strong sites ($F(2,622) = 29.34$, $p<.001$); the Newman-Keuls procedure for ad hoc comparisons showed a significant difference ($p<.05$) among all three classes. Linear regression analysis was performed relating the matrix means to cleavage strengths (numerically expressed as 1 to 5; see above) for the 65 cleavage sites; the correlation coefficient ($r=0.331$) was significant ($F(1,63) = 7.77$ $p<.01$). The linear regression equation obtained from this analysis was:
where \( s \) is the cleavage strength of a topoisomerase II site and \( m \) is its matrix mean score. The slope of this graph is fairly small and examination of the matrix means for sites of cleavage strengths 1 and 2 showed that average of the former is actually larger than that of the latter (data not shown); for this reason, the matrix mean was not used to predict cleavage strengths.

In comparison to the matrix mean model, linear regression relating unique scores and cleavage strength of the 65 sites shows a correlation coefficient which is numerically larger \((r=0.345 \text{ and } F(1,63) = 8.98, p<.01)\); however the correlation coefficient for the unique score is not greater than for the matrix mean by a statistically significant amount \((p \text{ is greater than .05})\). The linear relationship between unique score and cleavage strength is shown in the graph in Figure 20; the means for each class of cleavage strength are plotted and standard deviations are given in the legend. The equation for the regression line is:

\[
U = 0.55(s) + 6.32
\]

where \( U \) is the unique score and \( s \) is the strength of the cleavage site. The regression slope is 0.55 and the intercept is 6.32; the latter value is reasonable because
Figure 20. Topoisomerase II cleavage strength correlates significantly with the unique score. The 65 cleavage sites (in the absence of m-AMSA) were arranged by classes of relative cleavage strength (1-5) and for each class, the average unique scores were plotted versus the cleavage strength; the points are shown as circles. The actual values, accompanied by standard deviations are shown below:

<table>
<thead>
<tr>
<th>Relative Cleavage Strength</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Unique Score</td>
<td>7.0</td>
<td>7.4</td>
<td>7.8</td>
<td>8.7</td>
<td>9.5</td>
</tr>
<tr>
<td>Unique Std. Deviation</td>
<td>1.4</td>
<td>1.7</td>
<td>1.3</td>
<td>1.1</td>
<td>0.5</td>
</tr>
</tbody>
</table>
88% of the 65 cleavage sites have a unique score of at least 6 (these same sites have an average score of 8). Note, most sites that score less than a value of 6 are not cleaved (i.e., cleavage strength = 0). Given the validity and reliability of the consensus sequence information models for analysis of topo II sites, they have since been applied to the behavior of topo II towards other DNA sequences.

The topo II consensus sequence can accommodate significant homology with alternating purine-pyrimidine sequences. In fact, these repeating sequences were found to be hotspots for topo II cleavage (Spitzner et al, 1990). Examination of these sequences by the matrix mean model proved that the cleavage sites and strengths are proportional to the homology scores. Table 11A shows a sequence cloned from the upstream region of the human \( \beta \)-globin gene; this sequence contains a 54 base alternating purine-pyrimidine (RY) sequence. The topo II sites identified in the presence or absence of \( \text{m-AMSA} \) on this DNA fragment are indicated on the sequence. In Table 11B, portions of the sequence are shown with the matrix mean score for each potential site; clearly, the sites selected by topo II are those with the highest matrix mean scores, showing the validity of the model for predicting the behavior of the enzyme on an unusually reactive substrate. Other RY repeats examined also were cleaved by topo II, at sites predicted by the matrix mean (Spitzner et al, 1990).
Table 11. Recognition sites and homology scores in pRYG

**Recognition sites defined by cleavage and primer extension**

<table>
<thead>
<tr>
<th>Flanking sequence not cleaved by topo II</th>
<th>Recognition sites defined by cleavage and primer extension*</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGATTGGATTTACCTCTGTAAGAAAAAGAAAAAAA</td>
<td>ACTAAACCTATTTTGGAAAGACCATCTTTTTTTTTTTTTTTTT</td>
</tr>
<tr>
<td>ATATAATATGATATATATAGATATATACACATACATA</td>
<td>TATATATATATATATATATATATATATGTGTATGTAT</td>
</tr>
<tr>
<td>TACATATATATGCATTTGTTTGTGTTTTTCTTAAT</td>
<td>ATGTATATATACGTAAGTAAGAAAAAAAGAAAAA</td>
</tr>
</tbody>
</table>

**Homology scores in cleavage and non-cleavage (flanking) sequences of pRYG**

<table>
<thead>
<tr>
<th>Matrix Mean values:</th>
<th>Flanking sequence not cleaved by topo II (from above)</th>
</tr>
</thead>
<tbody>
<tr>
<td>259 220 248 279 168 266 245 158 242 235 209 243 262 266 240 226 248 275 255</td>
<td></td>
</tr>
<tr>
<td>CCTTCTGTAAAGAAAAGAAA</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Matrix Mean values:</th>
<th>pRYG sequences cleaved by topo II (from above)</th>
</tr>
</thead>
<tbody>
<tr>
<td>182 302 186 359 189 347 183 265 187 366 193 330 175 323 107 344 191 307 218 282</td>
<td></td>
</tr>
<tr>
<td>CATIACIATATIACIATIATIATIATIAT</td>
<td></td>
</tr>
</tbody>
</table>

*The sequence for the RY repeat of pRYG is given. Sites labeled "am" are cleavages mapped in the presence of m-AMSA; the boxed sites are those mapped in the absence of drug (from Figs. 2A and 2B). The flanking sequence in brackets represents a non-RY stretch that lacks topo II sites.

*The top strand of the non-RY sequence (in brackets in the top sequence) and an RY sequence containing a number of topo II cleavage sites were analyzed for matrix mean values. Cleavage sites are marked by vertical lines.*
DISCUSSION

Consensus methods make quantitative predictions of topo II sites.

The homology of a sequence to the topo II consensus sequence derived was shown to be a good indicator of whether it would be cleaved by topo II, but was not as reliable for predicting the frequency (or strength) of cleavage at the site. Subsequently, it was determined that if sequences are instead compared to a matrix of the base proportions of consensus positions, the matrix mean score (the average base proportion score for both strands of a site) correlated well with cleavage sites; in addition, the unique score (the number of consensus sequence matches on either strand) also correlated well with the strength of cleavage. The matrix mean and the unique score made accurate quantitative predictions of cleavage efficiency for topo II sites identified in the presence or absence of m-AMSA. To demonstrate the generality of the findings, a new independent data set of 65 topo II sites was compiled from cleavage reactions without the use of drugs. These sequences, ranging from weak to strong sites, were compared
directly with the original \textit{m}-AMSA induced sites. The results are shown in Table 10; they indicate that the new sites match the consensus sequence and matrix about as well as the sites used to derive them, demonstrating the validity of the consensus sequence. Clearly, the consensus sequence did characterize topo II sites and was not dependent on the peculiarities of one data set or the use of \textit{m}-AMSA. Thus, it was appropriate to do detailed analyses of topo II sites using the original consensus sequence information.

When the matrix means of the 65 topo II sites were compared to those of nonsites, analysis of variance showed a statistical significance (\(p<.001\)) between the sets, implying that the matrix mean discriminated well between sites and nonsites. A model that used both the matrix mean and the unique scores for a sequence was shown to be a good predictor of cleavage sites. Linear regression analysis showed that both the unique score and the matrix mean correlated well (\(p<.01\)) with the strength of cleavage. Together, these data suggest that the matrix mean and unique score models are useful for both qualitative and quantitative evaluation of topo II sites.

**Modelling the analysis methods**

The molecular basis for the unique score is not known; such a model has not been described previously. It is
likely to involve the concerted action of the homodimer; it has been shown that the cleavage events on the two strands are not simultaneous (Muller et al, 1988). It may be that the events of DNA binding and DNA cleavage by topo II have separable sequence requirements, where certain elements are required for binding by one monomer, while recognition elements required for cleavage could then be supplied by either monomer. Since there is a requirement for homology to the consensus sequence on both strands for efficient cleavage, it is may be that, prior to cleavage, each topo II monomer must be able to bind the substrate with a minimal binding affinity, then certain enzyme-DNA contacts must be made to allow the conformational changes involved in cleavage. The cleavage efficiency would then be related to the chances that proper interactions are made between either or both monomers and the necessary nucleotides; the monomers might complement each other such that binding energies needed for the conformational changes can be made in combinations of bases in contact with the two subunits.

The matrix mean model is more easily interpreted. This is based on a model proposed by Berg and von Hippel (see their 1989 review), which suggests that frequency of occurrence of a base in a matrix is proportional to its contribution to the binding energy of the protein-DNA interaction. Therefore, the matrix mean, which is the base proportions for a binding site (for both strands) would be
directly related to the binding energy of that site for its interaction with a topo II homodimer. Thus a site with a greater matrix mean score would be expected to have a higher binding affinity for topo II and would subsequently have a greater probability of being cleaved by the enzyme.

Most likely, the molecular interactions between topo II and DNA are approximated by elements from both of these models, but other factors that affect binding, such as DNA conformation and flexibility, will have to be included once such data becomes available (such as from specific mutagenesis analyses of sites). Until further information is obtained, however, these models should be valuable for predicting the actions of topo II in vivo as well as in vitro. For example, if the locations of potential cleavage sites were known it might be predicted that under conditions of high cellular levels of topo II (such as during mitosis; Heck et al, 1988), all sites would be utilized, but when topo II levels were low (as in G1 cells), only the strong sites (high matrix mean and unique scores) would be utilized.
Chapter V

Conservation of topoisomerase II recognition elements.

INTRODUCTION

In order to better understand the interaction between topo II and DNA, Sander and Hsieh (1985) sequenced a number of Drosophila topo II cleavage sites and derived a consensus sequence from those sites. After Cockerill and Garrard (1986) identified matrix associated regions (MARs) of DNA in a mouse immunoglobulin gene, they were interested in determining whether these sites were likely to interact with topo II (since it was shown that topo II is a major component of the nuclear scaffold, Earnshaw and Heck, 1985). The Drosophila topo II consensus sequence was the only such consensus identified, so they compared their MAR sequence to this and concluded that the MAR did contain topo II sites. The amino acid sequence had been determined for several organisms and there was only about 50% conservation between the enzymes (see above), but it was shown that they were functionally conserved in that the topo II from one organism could substitute for that of another. However, no studies
had been published comparing cleavage sites of topo II's from different species. Therefore, a consensus sequence was derived for the topo II of a higher eukaryote, chicken (Spitzner and Muller, 1988; see above), and this was compared to the fly consensus. The comparison of published consensus sequences showed little similarity between the two.

However, direct comparison of the two enzymes showed similar cleavage patterns for the fly and chicken enzymes; in addition, human topo II cleaved the same sites as the chicken enzyme. Therefore, a more extensive and quantitative analysis was done to compare topo II sites between chicken and fly. The results suggest that indeed there were conserved recognition elements between vertebrate and invertebrate topo II recognition elements.
MATERIALS AND METHODS

The reagents, cleavage reactions, analysis of data, and statistical analyses were as described in previous chapters.
RESULTS

Chicken and fly topo II sites are more related than their consensus sequences

In order to determine if the consensus sequence derived for chicken topo II can be applied to topoisomerases from other species, Chicken and Drosophila topo II cleavages were performed on the same kDNA fragment and analyzed on the sequencing gel shown in Figure 21. Enzyme concentrations were not high enough to reveal cleavages in the absence of m-AMSA; however, since the drug was used for both enzymes, the comparison is appropriate. The cleavage patterns are similar for the two enzymes. The chicken topo II sites matched the consensus sequence well: the two strongest sites showed homologies of 70 and 80% to the chicken consensus. However, the strongest Drosophila topo II site matched its published consensus sequence (Sander and Hsieh, 1985) only 46%, while sites of greater match (73% and 64%) 3 and 4 bases 3' of the strong site were not cleaved. Thus, while neither consensus sequence is perfect in its reliability, the chicken consensus sequence is more consistently accurate; one reason is that the fly topo II consensus was
Figure 21. Comparison of chicken and Drosophila topoisomerase II cleavage sites. Cleavage reactions were performed on a kDNA fragment (fragment number 1 of Table 1) using equal amounts of topo II activity (2 units) for the two enzymes. The gel (run from right to left) contains the following reactions: lanes 1, 3, and 5 contained the enzyme as indicated plus m-AMSA; lanes 2 and 4 contained Drosophila or chicken topo II and no drug; chemical sequencing reactions are in lanes 6 and 7. The DNA sequence is shown below the gel. There are a few high molecular weight fragments (lane 5) that were due to degradation of the parental DNA fragment and thus not topo II cleavage products.
drawn from a small data base of sequences with high A+T content, so that it makes good predictions on test sequences of this nature, but is not reliable for predicting sites on DNA of other base content. Overall, it is apparent that the site selections between the two enzymes have more in common than the consensus sequences would suggest (the fly consensus sequence is shown in Table 12).

Figure 22 shows a comparison of chicken, Drosophila, and human topo II cleavage sites on three DNA fragments. Again, the chicken and fly enzymes demonstrate similar cleavage patterns, while the chicken and human topoisomerases appear to cleave at essentially identical sites. Thus, the results suggest a conservation of topo II recognition elements across highly divergent organisms, and a common recognition pattern for vertebrate topo II enzymes; therefore, human topo II sites can be predicted using the same methods and consensus information as for chicken topo II sites.

Analysis of Drosophila sites using the chicken consensus

Despite the obvious differences between the chicken and Drosophila topo II consensus sequences (shown in Table 12a), it was clear that the enzymes must share some recognition elements, and these common elements might be more obvious if sites are compared by their individual base proportions. A
Figure 22. Comparison of topoisomerase II cleavages from different organisms. Different labeled fragments were subjected to cleavage reactions with no enzyme (dash), chicken enzyme (C), Drosophila enzyme (D), or human topo II (H). Gels I, II, and III contained, respectively, fragments 13, 14, and 1 from Table 1. Cleavage products were resolved on a 6% sequencing gel.
recent report by Lee, Sander, and Hsieh (1989) showed a set of 7 weak, 5 medium strength, and 10 strong *Drosophila* in vitro cleavage sites. Analysis of a strong site which showed a substantial DNase I footprint in fact also demonstrated a good match to the chicken topo II consensus information. The site scored the largest matrix mean value on the sequence (0.333), and also the greatest combined match (13), both of which are characteristic of strong chicken topo II sites. In fact, the average matrix mean score of these 22 fly sites, utilizing the chicken topo II matrix, is 0.315, exactly the average match of the 71 chicken topo II sites to their own matrix. In addition, as shown in Table 12B, the combined match and unique scores are proportional to the strength of these fly topo II cleavages. In fact, the combined match scores of these sites are about the same as for the no drug chicken topo II sites (see Chapter V). Therefore, it is apparent that the chicken and fly enzymes must have recognition elements in common, and the chicken topo II consensus information is useful for analysis of *Drosophila* topo II sites; however, the reverse is not true, that is, the *Drosophila* topo II consensus sequence does not discriminate well between sites and nonsites for the chicken enzyme (data not shown). If instead, the matrix of base proportions used to generate the fly consensus sequence is used to test the discrimination between chicken topo II sites and nonsites, the result is
Table 12. Comparison of Drosophila topo II sites with the chicken topo II consensus sequence information.

A. Comparison of consensus sequences (W is A or T, K is G or T).

Drosophila topo II consensus sequence:

5' GTNWAY*ATTNATNG 3' top strand
5' CNNAT*NAATRTWNAC 3' bottom strand
5' RNYNCCNGY*NGKTNYN 3' chicken topo II consensus sequence

B. Correlation between cleavage strength of fly topo II sites and average match to chicken topo II consensus sequence.

<table>
<thead>
<tr>
<th>Cleavage Strength</th>
<th>Combined Match</th>
<th>Unique Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>weak</td>
<td>10.1</td>
<td>6.0</td>
</tr>
<tr>
<td>medium</td>
<td>10.4</td>
<td>6.2</td>
</tr>
<tr>
<td>strong</td>
<td>10.9</td>
<td>6.5</td>
</tr>
</tbody>
</table>
that, on the whole, there was notable discrimination. Analysis of variance on fly matrix mean scores comparing sites with nonsites showed a highly significant correlation (p < .001).

Compilation of a new fly data base and analysis of topo II sites

One potential reason that the observed enzymatic sites for the two species are much more similar than the consensus sequences or the matrices are is because the Drosophila consensus was derived from a smaller sized sample of sequences, and also because the alignment procedures necessary before a consensus sequence can be determined (see Chapter IV) were conducted differently. The general problem of a consistent procedure for aligning sequences where common sequence elements might be found on either strand was solved by creating the redundancy maximization method. Since a consensus sequence is supposed to be, by definition, the sequence showing the maximum base information common to all the sites in the data base, and because redundancy is the measure of such information, then when the redundancy of a set of sequences has been maximized, the resulting consensus sequence will contain the most base information. In order to maximize the redundancy of a data set, the first thing that must be determined is the range of base
positions to evaluate; in other words, should the window examined be 10 bases on either side of the cleavage site, or 100 bases, or somewhere in between? This affects the results because the more nonessential (background noise) bases included, the greater the potential for aligning due to capitalizing on chance. For example, if a consensus sequence has 20 bases at a 95% confidence level, then one of those bases (5%) is probably there due to chance and not because it is a recognition element. Nuclease protection experiments have suggested that topo II makes contact with the DNA helix within a region 15 bases 5' of the top strand and bottom strand cleavage sites (Lee et al, 1989). Therefore, this window of 34 bases from -15 to +19 (+19 is -15 on the bottom strand) relative to the cleavage site was chosen.

In order to maximize the redundancy of a set of sequences, individual sequences are selected in random order and added to the previous sequences in either orientation; the alignment which gives a greater redundancy score is kept, and the next sequence is examined. To avoid any biases that might be introduced by the order in which the sequences are examined, the whole process is repeated many times (in different random orders) and then the alignment which gave the maximum redundancy score is used. The base proportions of this information-optimized data base can then be examined for derivation of a consensus sequence, or the
base proportion matrix can be used directly for further analysis (so that there are no arbitrary decisions made about handling the data). This analysis showed that, in fact, the sequences originally used to derive the *Drosophila* topo II consensus sequence (Sander and Hsieh, 1985) were not optimally aligned.

However, to create a more general fly topo II consensus sequence, a larger sample size was desirable; a set of 77 *Drosophila* topo II cleavage sites was compiled (Sander and Hsieh, 1985; Lee et al., 1989; Hsieh, personal communication) and subjected to the redundancy maximization procedure over the 34 base window described above. The base proportions of this aligned fly topo II data base are shown in Table 13. This matrix is useful for discriminating between sites and nonsites for either species of topo II (data not shown).

**Compilation and analysis of a multi-species data base**

The results suggest that the two topoisomerases do not cleave the identical sites, but that the recognition elements are still highly conserved between the two. Therefore, much of the apparent difference in consensus sequences may be due to the differences in base composition of the sequences analyzed for sites (the fly sequences were 70% A+T, while the chicken sequences had roughly equal proportions of all bases). A test of the similarity of topo
Table 13. Base proportions of the aligned *Drosophila* topo II data set.

<table>
<thead>
<tr>
<th>Pos</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>Percent</th>
<th>Redund.</th>
</tr>
</thead>
<tbody>
<tr>
<td>-15</td>
<td>0.387</td>
<td>0.093</td>
<td>0.240</td>
<td>0.280</td>
<td>7.11</td>
<td></td>
</tr>
<tr>
<td>-14</td>
<td>0.427</td>
<td>0.173</td>
<td>0.093</td>
<td>0.307</td>
<td>9.76</td>
<td></td>
</tr>
<tr>
<td>-13</td>
<td>0.342</td>
<td>0.132</td>
<td>0.118</td>
<td>0.408</td>
<td>9.67</td>
<td></td>
</tr>
<tr>
<td>-12</td>
<td>0.382</td>
<td>0.119</td>
<td>0.211</td>
<td>0.289</td>
<td>5.71</td>
<td></td>
</tr>
<tr>
<td>-11</td>
<td>0.338</td>
<td>0.143</td>
<td>0.182</td>
<td>0.338</td>
<td>4.70</td>
<td></td>
</tr>
<tr>
<td>-10</td>
<td>0.468</td>
<td>0.169</td>
<td>0.104</td>
<td>0.280</td>
<td>10.47</td>
<td></td>
</tr>
<tr>
<td>-9</td>
<td>0.403</td>
<td>0.169</td>
<td>0.052</td>
<td>0.377</td>
<td>14.30</td>
<td></td>
</tr>
<tr>
<td>-8</td>
<td>0.299</td>
<td>0.130</td>
<td>0.182</td>
<td>0.390</td>
<td>5.99</td>
<td></td>
</tr>
<tr>
<td>-7</td>
<td>0.156</td>
<td>0.156</td>
<td>0.117</td>
<td>0.571</td>
<td>17.04</td>
<td></td>
</tr>
<tr>
<td>-6</td>
<td>0.506</td>
<td>0.130</td>
<td>0.221</td>
<td>0.143</td>
<td>11.91</td>
<td></td>
</tr>
<tr>
<td>-5</td>
<td>0.143</td>
<td>0.338</td>
<td>0.143</td>
<td>0.377</td>
<td>6.92</td>
<td></td>
</tr>
<tr>
<td>-4</td>
<td>0.506</td>
<td>0.156</td>
<td>0.130</td>
<td>0.208</td>
<td>11.58</td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td>0.377</td>
<td>0.039</td>
<td>0.039</td>
<td>0.545</td>
<td>31.38</td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td>0.675</td>
<td>0.052</td>
<td>0.117</td>
<td>0.156</td>
<td>30.80</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>0.000</td>
<td>0.364</td>
<td>0.117</td>
<td>0.519</td>
<td>30.82</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.571</td>
<td>0.104</td>
<td>0.104</td>
<td>0.221</td>
<td>18.93</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.260</td>
<td>0.052</td>
<td>0.130</td>
<td>0.558</td>
<td>21.07</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.558</td>
<td>0.078</td>
<td>0.026</td>
<td>0.338</td>
<td>28.90</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.143</td>
<td>0.247</td>
<td>0.169</td>
<td>0.442</td>
<td>7.34</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.416</td>
<td>0.117</td>
<td>0.312</td>
<td>0.156</td>
<td>8.47</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.273</td>
<td>0.143</td>
<td>0.013</td>
<td>0.571</td>
<td>27.25</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.494</td>
<td>0.065</td>
<td>0.117</td>
<td>0.325</td>
<td>17.61</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.104</td>
<td>0.390</td>
<td>0.000</td>
<td>0.506</td>
<td>31.68</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.416</td>
<td>0.026</td>
<td>0.442</td>
<td>0.117</td>
<td>22.70</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.117</td>
<td>0.364</td>
<td>0.052</td>
<td>0.468</td>
<td>18.64</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.442</td>
<td>0.130</td>
<td>0.130</td>
<td>0.299</td>
<td>9.68</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.299</td>
<td>0.143</td>
<td>0.169</td>
<td>0.390</td>
<td>5.76</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.429</td>
<td>0.182</td>
<td>0.247</td>
<td>0.143</td>
<td>6.49</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.247</td>
<td>0.247</td>
<td>0.156</td>
<td>0.381</td>
<td>2.78</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.558</td>
<td>0.104</td>
<td>0.104</td>
<td>0.234</td>
<td>18.08</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.348</td>
<td>0.101</td>
<td>0.159</td>
<td>0.391</td>
<td>9.16</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.319</td>
<td>0.087</td>
<td>0.246</td>
<td>0.348</td>
<td>7.00</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.290</td>
<td>0.261</td>
<td>0.058</td>
<td>0.391</td>
<td>10.45</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.536</td>
<td>0.101</td>
<td>0.159</td>
<td>0.203</td>
<td>14.69</td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>0.359</td>
<td>0.156</td>
<td>0.143</td>
<td>0.342</td>
<td>6.04</td>
<td></td>
</tr>
<tr>
<td>Mean redundancy per position</td>
<td>14.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The proportions of each base are shown, by position relative to the cleavage site; the redundancy at each position is also indicated.
II recognition elements between the two species is to merge the sets. If there are major differences then the resulting interference will drastically decrease the amount of information in the data bases; however, if there are some base preferences common to both enzymes then this will be apparent, as these positions will still possess significant levels of information.

The data bases of 77 fly sites and 118 chicken sites (the 71 and 65 sequence sets were merged, and the duplicates, sites that were found in the presence or absence of m-AMSA, were eliminated) were combined into one composite data set, and the sequences in this combined set were optimally aligned by the redundancy maximization method. The redundancy of the aligned combined set is shown graphically by position in Figure 23, and the resulting base proportion matrix is shown in Table 14. Clearly, substantial information remains after the data bases were combined, showing that there are recognition elements common to both the fly and the chicken (and therefore, also the human) topoisomerase II enzymes. The total base information is lower than for just the fly sequences, but is higher than for the set of 65 no drug chicken sites (data not shown).

In addition, a goodness of fit test at 99% confidence (using the chi square equation with the matrix values versus either random base frequencies or the average base frequencies from the data base) confirmed that many positions contained base
Figure 23. Redundancy of composite data set, by position relative to the cleavage site. The redundancy values for the composite data base (from Table 14) are graphed by position relative to the cleavage site.
Table 14. Base proportions of the composite (chicken and fly) pooled data set after alignment.

<table>
<thead>
<tr>
<th>Pos</th>
<th>Proportions</th>
<th>Percent</th>
<th>Redund.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>-15</td>
<td>0.293</td>
<td>0.173</td>
<td>0.188</td>
</tr>
<tr>
<td>-14</td>
<td>0.267</td>
<td>0.120</td>
<td>0.241</td>
</tr>
<tr>
<td>-13</td>
<td>0.134</td>
<td>0.237</td>
<td>0.237</td>
</tr>
<tr>
<td>-12</td>
<td>0.268</td>
<td>0.149</td>
<td>0.227</td>
</tr>
<tr>
<td>-11</td>
<td>0.227</td>
<td>0.180</td>
<td>0.206</td>
</tr>
<tr>
<td>-10</td>
<td>0.392</td>
<td>0.165</td>
<td>0.263</td>
</tr>
<tr>
<td>-9</td>
<td>0.253</td>
<td>0.191</td>
<td>0.170</td>
</tr>
<tr>
<td>-8</td>
<td>0.231</td>
<td>0.169</td>
<td>0.195</td>
</tr>
<tr>
<td>-7</td>
<td>0.179</td>
<td>0.133</td>
<td>0.303</td>
</tr>
<tr>
<td>-6</td>
<td>0.297</td>
<td>0.123</td>
<td>0.267</td>
</tr>
<tr>
<td>-5</td>
<td>0.133</td>
<td>0.359</td>
<td>0.138</td>
</tr>
<tr>
<td>-4</td>
<td>0.385</td>
<td>0.067</td>
<td>0.344</td>
</tr>
<tr>
<td>-3</td>
<td>0.251</td>
<td>0.072</td>
<td>0.190</td>
</tr>
<tr>
<td>-2</td>
<td>0.385</td>
<td>0.097</td>
<td>0.395</td>
</tr>
<tr>
<td>-1</td>
<td>0.113</td>
<td>0.323</td>
<td>0.118</td>
</tr>
<tr>
<td>1</td>
<td>0.431</td>
<td>0.123</td>
<td>0.179</td>
</tr>
<tr>
<td>2</td>
<td>0.246</td>
<td>0.087</td>
<td>0.344</td>
</tr>
<tr>
<td>3</td>
<td>0.297</td>
<td>0.077</td>
<td>0.267</td>
</tr>
<tr>
<td>4</td>
<td>0.174</td>
<td>0.118</td>
<td>0.144</td>
</tr>
<tr>
<td>5</td>
<td>0.390</td>
<td>0.205</td>
<td>0.262</td>
</tr>
<tr>
<td>6</td>
<td>0.272</td>
<td>0.164</td>
<td>0.067</td>
</tr>
<tr>
<td>7</td>
<td>0.328</td>
<td>0.138</td>
<td>0.185</td>
</tr>
<tr>
<td>8</td>
<td>0.128</td>
<td>0.267</td>
<td>0.138</td>
</tr>
<tr>
<td>9</td>
<td>0.231</td>
<td>0.133</td>
<td>0.390</td>
</tr>
<tr>
<td>10</td>
<td>0.164</td>
<td>0.179</td>
<td>0.185</td>
</tr>
<tr>
<td>11</td>
<td>0.236</td>
<td>0.149</td>
<td>0.215</td>
</tr>
<tr>
<td>12</td>
<td>0.221</td>
<td>0.205</td>
<td>0.236</td>
</tr>
<tr>
<td>13</td>
<td>0.251</td>
<td>0.179</td>
<td>0.297</td>
</tr>
<tr>
<td>14</td>
<td>0.277</td>
<td>0.195</td>
<td>0.195</td>
</tr>
<tr>
<td>15</td>
<td>0.287</td>
<td>0.154</td>
<td>0.215</td>
</tr>
<tr>
<td>16</td>
<td>0.215</td>
<td>0.161</td>
<td>0.242</td>
</tr>
<tr>
<td>17</td>
<td>0.200</td>
<td>0.200</td>
<td>0.254</td>
</tr>
<tr>
<td>18</td>
<td>0.201</td>
<td>0.207</td>
<td>0.283</td>
</tr>
<tr>
<td>19</td>
<td>0.321</td>
<td>0.245</td>
<td>0.174</td>
</tr>
<tr>
<td>Sum</td>
<td>0.255</td>
<td>0.169</td>
<td>0.228</td>
</tr>
<tr>
<td>Mean redundancy per position</td>
<td>6.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The proportions are shown, by position, for each nucleotide in the data base. The redundancy values are also shown.
proportions that were significantly different from background (data not shown).

If a consensus sequence is derived from the composite set base proportions (see Table 15), it can be seen that this has considerable homology to the chicken topo II consensus sequence. This new matrix of base proportions can be used to analyze chicken or *Drosophila* topo II sites with greater consistency than when the heterologous matrices were used (data not shown); however, as expected, it did not predict sites within a given data base as well as did the matrix constructed from that data base.
Table 15. Comparison of the composite data set consensus sequence with the chicken topoisomerase II consensus sequence.

Composite set consensus sequence\(^a\) (with cleavage at the \(\hat{\text{\,}}\)):

\[
\begin{align*}
\text{N} & \quad \text{N} & \quad \text{N} & \quad \text{N} & \quad \text{N} & \quad \text{C} & \quad \text{A} & \quad \text{A} & \quad \text{A} & \quad \text{C} & \quad \text{N} & \quad \text{N} & \quad \text{A} & \quad \text{T} & \quad \text{N} & \quad \text{T} & \quad \text{N} & \quad \text{C} \\
\text{T} & \quad \text{G} & \quad \text{T} & \quad \text{G} & \quad \text{T} & \quad \text{G} & \quad \text{T} & \quad \text{T} & \quad \text{T} & \quad \text{T} & \quad \text{T} & \quad \text{T} & \quad \text{T} & \quad \text{T} \\
-10 & \quad -8 & \quad -5 & \quad -2 & \quad -1 & \quad +2 & \quad +3 & \quad +4 & \quad +6 & \quad +8 \\
\text{A} & \quad \text{N} & \quad \text{C} & \quad \text{N} & \quad \text{N} & \quad \text{C} & \quad \text{N} & \quad \text{N} & \quad \text{G} & \quad \text{C} & \quad \text{N} & \quad \text{G} & \quad \text{G} & \quad \text{T} & \quad \text{N} & \quad \text{C} & \quad \text{N} & \quad \text{C} \\
\text{G} & \quad \text{T} & \quad \text{T} & \quad \text{T} & \quad \text{T} & \quad \text{T} & \quad \text{T} & \quad \text{T} & \quad \text{T} & \quad \text{T} & \quad \text{T} & \quad \text{T} & \quad \text{T} & \quad \text{T} & \quad \text{T}
\end{align*}
\]

Chicken topo II consensus sequence (above, 5' to 3')

\(^a\)Positions were included in the consensus sequence if the proportion of a single base was greater then 49% or, failing that, if the proportions of two bases combined to at least 70% at a position.
DISCUSSION

Direct comparison of DNA cleavage patterns for type II topoisomerases from chicken, Drosophila, and human sources showed that all three recognized similar sites, and that the two vertebrate enzymes cleaved essentially identical sites (Figures 21 and 22). It appeared from Figure 22 and from other data (not shown) that the recognition elements for human topo II were identical to those of the chicken enzyme. Furthermore, despite very different published consensus sequences (see Table 12), chicken and Drosophila topo II recognition elements must, in fact, have much in common. It was likely that some of the difference was due to differences in sequences analyzed (see above); since the sites selected by both enzymes in Figure 21 matched the chicken consensus sequence better than the fly, and because the former was derived from a greater variety of sequences, it was decided to examine fly sites using chicken consensus information.

The matrix mean was used because it does not require arbitrary decisions about determining whether a position is, for example, a guanine or a purine. Examination of 22 fly
sites recently identified by Lee et al (1989) showed that these sites had matrix mean values identical to those of chicken topo II sites; moreover, the combined and unique scores of the sites correlated well with their cleavage strengths (Table 12B). Analysis of variance showed that chicken enzyme matrix mean scores discriminated well between Drosophila topo II sites and nonsites (p<.001). The reverse was only somewhat true, that is, the fly site matrix did discriminate between chicken sites and nonsites, but not well.

Examination of the fly sites by redundancy analysis showed that they were not optimally aligned. To derive results with more generality, a larger fly topo II site data base was compiled. The sites were aligned by maximizing the redundancy; the resulting matrix was a better predictor of topo II sites from fly or chicken (data not shown). If the hypothesis is that there are certain elements conserved between chicken and fly topo II's, then combining the data sets would show consistent information at similar positions, while differences would result in a relative loss of information at other positions. The data bases were combined and all sites were aligned by maximizing the redundancy within the whole set. As the redundancy plot in Figure 23 shows, there remained a considerable amount of information in the new data set. A goodness of fit test at 99% confidence confirmed the significance of the information
at many positions in the combined data set. This suggested a conservation of recognition elements between fly and chicken topo II's (and therefore, also human topo II). If a consensus sequence was derived from this data base, it in fact resembled the chicken consensus (see Table 15), confirming that it was the more general result. The matrix from the combined data exhibited more accuracy in prediction of sites from individual data sets than did heterologous matrices, but the new matrix was not as reliable for predicting sites as were the homologous matrices. Thus, there is some conservation of topo II recognition elements between widely divergent species. However, there were also differences in site selection between the two topo II's; indeed, this might have been predicted by the fact that the protein coding regions are only about 50% conserved between topo II's of divergent species (see above). The results of further experiments will required to fully characterize the similarities and differences in topo II recognition elements between species.
INTRODUCTION

A by-product of the development of a means to separate large DNA molecules electrophoretically was the embedding of DNA in low melting temperature agarose to prevent DNA shearing (Schwarz and Cantor, 1984). High molecular weight DNA was shown to remain within the pores of the gel inserts and not diffuse out. However, smaller molecules, such as restriction enzymes, diffused freely within the inserts. This suggested the possibility that small molecules of DNA could diffuse out of the inserts. Kinetoplast DNA (kDNA) from trypanosome mitochondria consists of highly catenated networks with very large molecular weights, and only topo II can decatenate these networks and yield monomer circles (Marini et al, 1980). Therefore, the hypothesis was that if kDNA is embedded in low melt agarose inserts, it can not diffuse out; however, if topo II is active within the inserts, it can decatenate the networks and the monomer
circles could then diffuse out of the gel inserts. Indeed, it was shown that topo II can decatenate embedded kDNA, and the amount of released monomer is proportional to the topo II activity. It was also shown that if the kDNA is radioactively labeled, the amount of radioactivity that diffuses out is a measure of the topo II activity. Thus, the assay can be used to monitor topo II activities or identify topo II inhibitors.
MATERIALS AND METHODS

Low melting temperature agarose was from FMC (InCert genetic technology grade). 96 well microtiter plates were from Falcon. Other reagents were as described in previous chapters. Decatenation reactions and gel electrophoresis were also as described previously.
RESULTS

Topo II releases kDNA monomers from agarose inserts

The results of the experiment shown in Figure 24 revealed that topo II could decatenate kDNA that was embedded in low melt agarose inserts. Furthermore, most of the decatenated products diffused out of the inserts, while catenated kDNA remained in the insert (compare lanes 1-6, in which the insert was loaded onto the gel, with lanes 10-15, the supernatants from the reactions). In addition, the reaction appeared to reach complete decatenation after 4.5 hr incubation; however, the amount that diffuses out seems to reach its maximum after 2.25 hr. Lengthy incubation (15 and 22 hr) led to the production of linear kDNA molecules, contaminants which are not desirable because they are not strictly mediated by topo II, unlike the formation of relaxed monomer minicircles.

The next experiment (Figure 25) showed that the decatenation reactions could be carried out in 10 μL miniplugs in 96 well microtiter plates. Topo II decatenation products (relaxed and open circular kDNA minicircles) were observed after 30 min incubation and
Figure 24. Topo II activity in low melt agarose inserts. 5 μg kDNA were suspended in 55 μL TE and mixed with 55 μL 1% low melt agarose (previously melted, then equilibrated at 42°C) and a 100 μL rectangular insert was cast and allowed to solidify at 4°C. The insert was cut into 6 equal pieces and placed in microcentrifuge tubes to which 25 μL water and 5 μL 10X topo II reaction buffer were added. Then 5 μL topo II were added to lanes 2-6 and reactions were incubated at 30°C for the following times: lane 1, 8 hr; lane 2, 1 hr; lane 3, 2.25 hr; lane 4, 4.5 hr; lane 5, 15 hr; and lane 6, 22 hr. After incubation, the supernatants were removed from the inserts, mixed with 3 μL 10X loading dye, and loaded on lanes 10-15 (equivalent to the reactions of lanes 1-6, respectively) of a 1% agarose gel. 30 μL 1% sarkosyl, 0.5 M EDTA were added to the inserts and incubated 15 min at room temperature, then the inserts were removed and placed in the wells (lanes 1-6) of the agarose gel. 0.5% molten low melt agarose was poured around the inserts in the wells and allowed to harden at 4°C. Lanes 7, 8, and 9 were controls: decatenated kDNA, catenated kDNA, and linear kDNA, respectively (the mobility of these species is indicated). The gel was run in TAE buffer with 0.5 μg/mL ethidium bromide for 1 hr at 5 volts/cm.
Figure 25. Time dependence for release of decatenated kDNA from low melt agarose inserts. An insert mixture was made as in Figure 24, but instead was cast as 10 μL miniplugs in 9 wells of a 96 well test plate. 1.3 μL 10X topo II reaction buffer was added to each well, followed addition of topo II and incubation at 30°C as specified: lane 1, no topo II, 1 hr; lane 2, 1 μL topo II, 1 hr; lane 3, 2 μL topo II, 1 hr; lane 4, 4 μL topo II, 1 hr; lane 5, 6 μL topo II, 1 hr; lane 9, 2 μL topo II, 15 min; lane 10, 4 μL topo II, 15 min; lane 11, 2 μL topo II, 30 min; lane 12, 4 μL topo II, 30 min. After incubation, supernatants were removed, added to 5 μL stop dye, and loaded on a 1% agarose gel. Lanes 6 and 7 are decatenated kDNA markers (OC and Rel) and lane 8 is catenated kDNA. The gel was run in TAE with 0.5 μg/mL ethidium bromide at 5 volts/cm for 45 min.
peaked after 60 min incubation. It appeared that the topo II concentration was not a major factor in the rate of the reaction; the rate-limiting step is therefore likely to be diffusion of the DNA out of the agarose insert.

Whole trypanosomes (rather than purified kDNA) were embedded in inserts by the method of C. Cantor and the DNA was purified within the inserts. These inserts containing genomic and kinetoplast DNA were incubated for 7.5 hr with topo II, then the inserts and the supernatants were analyzed by gel electrophoresis (as shown in Figure 26). As with the purified kDNA, the decatenated products diffused into the supernatant (lanes 2-5), while high molecular weight DNA (genomic DNA and any catenated kDNA) remained in the inserts. The decatenation went equally well with 3, 5, or 10 μL topo II used, but the reaction with only 1 μL was not as efficient. Unlike topo II, the restriction enzyme XhoI released almost all of the DNA (genomic and kDNA alike) into the supernatant (lanes 6 and 14); this suggested that topo II was carrying out its normal enzymatic activity, and not cleaving DNA, which would lead to a release of a smear of different DNA sizes, as in lane 6.
Topo II decatenates kDNA from total *Crithidia fasciculata* DNA purified in agarose inserts. 1.2 x 10^9 cells were washed in PSG (75 mM sodium phosphate pH 8, 65 mM NaCl, 10% glucose) and resuspended in a total volume of 300 μL with PSG, then mixed with 300 μL molten 1% low melt agarose and cast as 100 μL inserts. After solidifying, the inserts were digested in 0.5 M EDTA pH 9.5, 1% sarkosyl, 1 mg/mL Proteinase K at 50°C for 48 hr, then washed (at 37°C) twice in TE with 1 mM PMSF for 24 hr and three times in TE for 36 hr (total). 1 insert was cut into 6 equal pieces and each was added to a microcentrifuge tube in which 50 μL reactions were run in topo II reaction buffer. Topo II was added to lanes 1-5 in the amounts of 0, 1, 3, 5, and 10 μL, respectively, and tubes were incubated 7.5 hr at 30°C; in lane 6, 10 units XhoI were added, followed by 7.5 hr incubation at 37°C. Then the supernatants were mixed with stop dye and loaded on a 1% agarose gel. The inserts after digestion were incubated 10 min in 10 μL stop dye then loaded (1-6, respectively) onto lanes 9-14 of the gel. Lanes 7 and 8 are decatenated (OC and Rel) kDNA and linear (Lin) kDNA markers.
Use of radiolabeled kDNA in gel release assay allows rapid large-scale screens of topo II activity

For analysis of topo II activities in a large number of samples, gel electrophoresis becomes prohibitive; only a few gels can be run in a single day. For this reason, the gel release assay with kDNA in agarose inserts was adapted to using $^3$H labeled kDNA and simply counting the radioactivity of decatenated DNA released into the supernatant. The low melt agarose concentration was lowered to 0.25% to allow more rapid diffusion of enzyme and decatenated products. As the results in Figure 27 indicate, addition of topo II to the inserts resulted in a substantial increase in the amount of radioactive DNA detectable in the supernatant (three times the background counts). The decatenation reaction of topo II was decreased significantly by inclusion of the topo II inhibitor m-AMSA, (a reduction in $^3$H counts of 50% after subtracting out the background radiation released). Lane 4 shows that other agents which degrade DNA will also release $^3$H into the supernatant. Thus, the gel release assay can be used to screen topo II inhibitors (resulting in a decrease of radioactivity released) and identify agents with DNase activities (they release $^3$H into the supernatant without requiring topo II); of course, a substance with both those activities might be missed in the screen.
Figure 27. Quantitative use of the gel release assay. Tritiated kDNA was cast in miniplugs as in Figure 25 (except with agarose at 0.25%) at 3000 cpm $^3$H per plug. Duplicate reactions 1-4 were run in a 96 well microtiter plate. 20 µL overlays containing topo II reaction buffer were added, with no enzyme in reaction 1, 2 µL topo II in reaction 2, 2 µL topo II and 1 µg m-AMSA in reaction 3, and 10 µg DNaseI in reaction 4. After 30 min incubation at 37°C, the supernatants were removed to scintillation vials and the miniplugs were washed with 100 µL TE which was added to the vials. The samples were then counted for $^3$H. The results were averaged for the duplicates and are shown as a bar graph.
DISCUSSION

It was demonstrated that if kDNA was embedded in low melt agarose, the catenated networks would not diffuse out. However, under appropriate reaction conditions, topo II could diffuse in and be catalytically active, leading to the release of decatenated kDNA monomers in the supernatant. A method for the large-scale screening of topo II activity was desired; therefore, gel electrophoresis becomes a limiting factor. To develop the assay for rapidity, low melt agarose inserts containing the kDNA were cast in 96 well microtiter plates, and the reactions were performed in the wells, followed by removal of the supernatant. An adaptation to relieve the need for gel electrophoresis was to label the kDNA with $^3$H-thymine in the trypanosome culture. Then the reactions could be run as above, but scintillation counting of samples replaced the need to run gels. As shown in Figure 27, radioactivity was released by this assay in proportion to the expected topo II activity, and the topo II inhibitor m-AMSA blocked the release of radiolabeled DNA from the gel insert. In addition, it was shown that the assay could be used without topo II to detect the action of a nuclease on the kDNA.
This gel release assay has been used for screening column chromatography fractions for topo II activity (not shown), and has also been shown to be valuable for rapid screening of inhibitors of topo II (data not shown). Whereas perhaps only 100 samples could be assayed for topo II activity by gel electrophoresis in one day, several hundred reactions could be run and analyzed by a single investigator in one day with the gel release assay, and the use of multipipetors could greatly speed the time to set up the reactions. Therefore, the gel release assay could prove valuable for screening large numbers of compounds for topo II inhibitors that might later be used for anticancer chemotherapy. Once initial results are obtained from this assay, further tests would need to be done to investigate the findings in detail.
SUMMARY

The primary goal of this research project was to characterize the recognition elements for the cleavage of DNA by type II topoisomerases. To this end, topo II was purified from chicken erythrocytes in a novel rapid purification scheme. The enzyme preparation was shown to be a homogeneous band on an SDS-polyacrylamide gel by Coomassie Blue staining, and also by western blots using a specific antibody against topo II. The chicken erythrocyte topo II possessed the characteristics common to other eukaryotic type II enzymes examined. Monoclonal antibodies were generated against topo II and shown to be specific; however, while these antibodies were valuable for detecting topo II by western blots, they failed to localize topo II within cells by immunofluorescence. The major focus of the topo II research project was to formulate a model to both describe and predict the interaction between topo II and DNA sequences. In order to obtain a model general enough to be of value in describing sites not in the test sample, topo II cleavage reactions were performed on a large number of diverse DNA sequences. The sequences of 71 strong cleavage
sites were used to derive a consensus sequence. The consensus sequence was highly degenerate, but this was expected as topo II binds to many different sequences; in particular, the value of the consensus was demonstrated by its use to predict cleavage sites in untested DNA. Detailed analyses also suggested that site recognition by topo II involved specific interactions between enzyme subunits and both strands of a DNA cleavage site.

The conventional consensus sequence model provided accuracy of site prediction, but not so much of quantitation; that is, the relative frequencies of cleavage at sites were not associated well with consensus matches. Additional characterizations of topo II sites were made to derive a quantitative model. One method was the unique score, which represents the consensus matches found on either strand of a potential site. A more general method was the matrix mean model. This model is based on the assumption that if test sites are generated from an unbiased sample, the base proportions observed in sites will represent the true preferences of the enzyme for those positions. Such a model was desirable because it avoids the arbitrary nature of consensus sequences. The matrix mean proved to be of value for quantitative analyses of sites.

The generality of the models was demonstrated by showing that they were accurate in describing an independent data set. In addition, it was shown that while the drug m-
AMSA can influence the selection of topo II sites, the recognition patterns are not significantly different. The consensus information was also found to accurately predict the great reactivity of topo II to alternating purine-pyrimidine sequences. Finally, chicken topo II sites were compared with those of other species. Human topo II cleaved the same sites as the chicken enzyme, suggesting that the chicken consensus sequence could be applied to all vertebrate topo II's. A consensus sequence identified for the Drosophila enzyme (Sander and Hsieh, 1985) was very different from the chicken consensus; however, the two enzymes clearly demonstrated similarities in site selection. An analysis of chicken and fly site data bases by the matrix methods that were developed here indicated that, in fact, there were recognition elements common to the two topoisomerases. The final set of experiments presented shows the development of a novel rapid and quantitative assay for topo II activity. The assay is based on the observation that topo II could decatenate kDNA networks that were embedded in low melt agarose. The assay is specific for topo II, and use of radiolabeled kDNA obviates the need for electrophoretic separations; levels of topo II activity can be measured by scintillation counting of reaction supernatants from 96 well microtiter plates. This a major simplification for large-scale topo II screens, both to assay for enzyme activity (from column fractions, for
example) and to screen for inhibitors of topo II (which are of enormous interest because they have great potential for anticancer chemotherapy). In conclusion, this body of work describes a number of new and significant observations which should aid in eventually understanding the interactions between type II topoisomerases and DNA.
LIST OF REFERENCES


