doi:10.1016/S0022-2836(03)00243-2 *J. Mol. Biol.* (2003) **328**, 9–32

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Assessment by Molecular Dynamics Simulations of the Structural Determinants of DNA-binding Specificity for Transcription Factor Sp1

Esther Marco, Raquel García-Nieto and Federico Gago*

Departamento de Farmacología
Universidad de Alcalá
Alcalá de Henares, E-28871
Madrid, Spain

The DNA-binding domain (DBD) of the ubiquituous transcription factor Sp1 consists of three consecutive zinc fingers that recognize a number of nucleotide sequences different from, but related to and sometimes overlapping, those recognized by the structurally better characterized early growth response protein 1 (EGR1, also known as Zif268, Krox-24, and NGFI-A). The accepted consensus binding sequence for Sp1 is usually defined by the asymmetric hexanucleotide core GGGCGG but this sequence does not include, among others, the GAG (= CTC) repeat that constitutes a high-affinity site for Sp1 binding to the *wt1* promoter. Since no 3D structure of the whole DBD of Sp1 is available, either alone or in complex with DNA, a homology-based model was built and its interaction with two DNA 14-mers was studied using nanosecond molecular dynamics simulations in the presence of explicit water molecules. These oligonucleotides represent Sp1 target sites that are present in the promoters of the *mdr1* and *wt1* genes. For comparative purposes and validation of the protocol, the complex between the DBD of EGR1 and its DNA target site within the proximal *mdr1* promoter was simulated under the same conditions. Some water molecules were seen to play an important role in recognition and stabilization of the protein–DNA complexes. Our results, which are supported by the available experimental evidence, suggest that the accuracy in the prediction of putative Sp1-binding sites can be improved by interpreting a set of rules, which are a blend of both stringency and tolerance, for the juxtaposed triplet subsites to which each zinc finger binds. Our approach can be extrapolated to WT1 and other related natural or artificial zinc-finger-containing DNA-binding proteins and may aid in the assignment of particular DNA stretches as allowed or disallowed-binding sites.

Keywords: transcription factors; zinc fingers; EGR1/Zif268/Krox-24/NGFI-A; Sp1; molecular dynamics simulations

*Corresponding author

Introduction

Sp1 is a representative member of an expanding family of mammalian Sp/XKLF zinc finger proteins that regulate the expression of more than 1000 different genes involved in cell proliferation, differentiation and apoptosis through their binding to G/C-rich *cis*-regulatory sequences and subsequent interaction with the basal transcription machinery.[1–3] These transcription factors, which can behave as activators or repressors, are classified into different subfamilies (e.g. Sp, BTEB, KLF, CPBP, and TIEG) and the sequences to which they bind are present in the promoters of genes encoding, for example, cell-cycle regulators (p15/INK4B, p21/WAF1, p27/KIP1), MAP kinases, regulatory GTPases (Ha-Ras), histones, enzymes involved in DNA synthesis (thymidine kinase, dihydrofolate reductase), growth factors (TGF-β1, PDGF), and growth factor receptors (insulin receptor, insulin-like growth receptor).

Abbreviations used: DBD, DNA-binding domain; EGR1-f1, EGR1-f2 and EGR1-f3, fingers 1, 2 and 3 of EGR1; ET743, ecteinascidin 743; MDR, multidrug resistance; PDB, Protein Data Bank; Sp1-f1, Sp1-f2 and Sp1-f3, fingers 1, 2 and 3 of Sp1; TPA, 12-*O*-tetradecanoylphorbol-13-acetate; MEP, molecular electrostatic potential; nt, nucleotide.

E-mail address of the corresponding author: federico.gago@uah.es

0022-2836/03/$ - see front matter

```
                534                                                    591
sp1_human        KKQHICHIQGCGKVYGKTSHLRAHLRWHTGERPFMCTWSYCGKRFTRSDELQRHKRTH
sp2_human        KKKHVCHIPDCGKTFRKTSLLRAHVRLHTGERPFVCNWFFCGKRFTRSDELQRHARTH
sp3_human        KKQHICHIPGCGKVYGKTSHLRAHLRWHSGERPFVCNWMYCGKRFTRSDELQRHRRTH
sp4_human        KKQHICHIEGCGKVYGKTSHLRAHLRWHTGERPFICNWMFCGKRFTRSDELQRHRRTH
wt1_human        KRPFMCAYPGCNKRYFKLSHLQMHSRKHTGEKPYQCDFKDCERRFSRSDQLKRHQRRH
egr1_human       ERPYACPVESCDRRFSRSDELTRHIRIHTGQKPFQCR--ICMRNFSRSDHLTTHIRTH
                102                                                    157
                     *    *            *   *       *    *           *   *


                592                          620
sp1_human        TGEKKFACPECPKRFMRSDHLSKHIKTHQ
sp2_human        TGDKRFECAQCQKRFMRSDHLTKHYKTHL
sp3_human        TGEKKFVCPECSKRFMRSDHLAKHIKTHQ
sp4_human        TGEKRFECPECSKRFMRSDHLSKHVKTHQ
wt1_human        TGVKPFQCKTCQRKFSRSDHLKTHTRTHTGKTSEKPFSCRWPSCQKKFARSDELVRHHNMHQR
egr1_human       TGEKPFACDICGRKFARSDERKRHTKIHL
                158                          186
                       *  *            *   *          *    *           *   *
```

Figure 1. Amino acid sequences of the zinc finger domains of relevant Sp1 family members, EGR1, and WT1. Asterisks designate cysteine and histidine residues that are involved in zinc ion coordination. Amino acid residues in the recognition α-helix (from −1 to + 6) that make contacts with DNA bases are underlined. Sp1 and EGR1 sequences are numbered.

The DNA-binding domain (DBD) of Sp1-like proteins, which is located at the C terminus and comprises three consecutive zinc fingers (Figure 1), is believed to make use of a well-studied protein-DNA recognition code that is based on a relatively simple set of contacts.[4,5] Sp1 was initially reported to bind to the asymmetric hexanucleotide core GGGCGG, with a consensus sequence 5′-(G/T)GGGCGG(G/A)(G/A)(C/T)-3′ (Figure 2).[6]

Sp1 bound to some of these GC-rich positive regulatory elements is known to be displaced in a dose-dependent manner by elevated amounts of early growth response protein 1 (EGR1, also known as Zif268, Krox-24, and NGFI-A), which results in decreased Sp1-dependent transcription.[7] The consensus sequence that is usually accepted for an EGR1 high-affinity-binding site is 5′-GCG(G/T)GGGCG-3′.[8] Therefore, EGR1 and the ubiquitously expressed Sp1 recognize different, but related, nucleotide sequences that can be partially overlapped,[9] in which case their binding is mutually exclusive.

The *sp1* gene has been mapped to 12q13.1,[10] a chromosomal region commonly involved in rearrangements in soft tissue tumours, and the chromosomal location of *egr1* is 5q23-31, a region that is often deleted in epithelial hyperproliferation and myeloid disorders.[11,12] For example, human HT-1080 fibrosarcoma cells are known to express little or no EGR1, and only mutant p53. In these cells, stable expression of exogenous EGR1 or overexpression of just the DBD of this transcription factor inhibits transformed growth in a dose-dependent manner and causes decreased tumorigenicity.[13] This effect can be inhibited[14] by expression of the Wilms' tumour suppressor (WT1), a known specific DNA-binding competitor possessing four zinc fingers in its DBD (Figure 1). WT1, which is found mutated in Wilms' tumour and other nephropathies,[15] binds to the "EGR1 consensus site", as well as other sites,[16] and may repress (e.g. *tgf-β1*) or stimulate (e.g. *bcl-2*) gene expression depending on promoter context. Since DNA binding of WT1, in turn, has been shown recently to be inhibited by association with members of the p53 family of tumor suppressors,[17] a complex pattern of cross-regulation, and important roles in cell growth, differentiation and programmed cell death are apparent for many of

Sequence	Source
TGG GCG GAG T	*sv40* (III, V)
GGG GAG TGG C	HIV-LTR (I)
TGG GCG GGA C	HIV-LTR (II)
GAG GCG TGG C	HIV-LTR (III)
TGG GCG GGG C	*hsv IE-3* (III, IV)
GGG GCG GGG C	*hsv IE-3* (V), *dhfr* (I, III), metallothionein IIa (I, II)

Sequence	Source
GGG GCG GAG C	*dhfr* (II, IV)
GGG GCG TGG G	human *mdr1*
GGG GCG GGG C	human *sp1*
GAG GAG GAG C	*wt1*
GAG GAG GGG G	*timp-2*
GAG GCG GAG C	human *TERT*

Figure 2. Examples of high-affinity-binding sites for Sp1.[6,18,29,56,67,97,98] Gaps reflect subsite triplet separations.

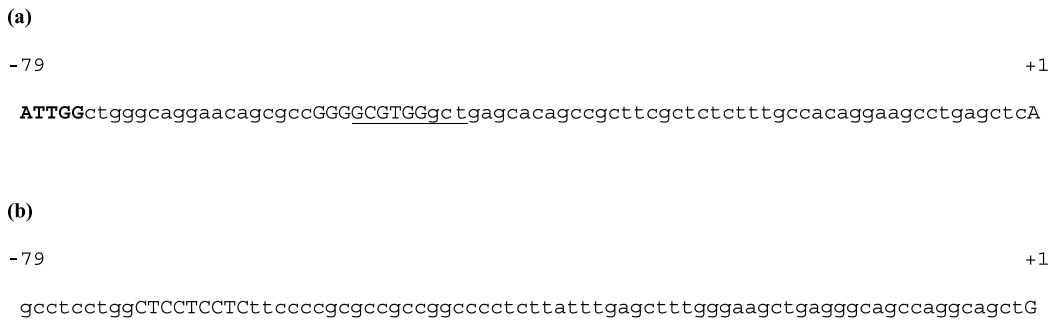

Figure 3. Nucleotide sequences of the 5′-flanking regions (from −79 to + 1) of (a) the human *WT1* promoter[18] showing the inverted Sp1-binding site (upper case), and (b) the human *mdr1* promoter[21] showing the Sp1-binding site (upper case) downstream of the inverted CCAAT element (bold type). Overlapping this site is the minimal nona-nucleotide-binding site for EGR1 (underlined). Numbering in both cases is relative to the major transcription start site (+1).

these transcription factors. As a further example of reciprocal modulation, the major enhancer element in the *wt1* gene is a single high-affinity Sp1 site located immediately upstream of the transcription start site (Figure 3).[18] Remarkably, this site consists of a $(CTC)_3$ repeat that apparently does not comply with the usually accepted consensus for Sp1 binding. In line with this observation are the reports that WT1 binds to $(TCC)_n$ tracts[19] and that EGR1 can bind with high affinity to sequences such as GAG(T/G)GGGAG, which includes no cytosine in the G-rich strand.[20] Therefore, the variability in binding sites for these transcription factors is greater than is generally acknowledged and not completely understood.

A well-studied example of this interplay between related transcription factors within a discrete DNA stretch takes place at the TATA-less promoter of the *mdr1* gene,[21] which encodes a highly conserved 180 kDa membrane P-glycoprotein that mediates the efflux of different xenobiotics from the cytoplasm and is found overexpressed in cancer cells exhibiting a multidrug resistance (MDR) phenotype.[22] In fact, *mdr1* expression can be activated rapidly in human tumors during the course of chemotherapy[23] and is usually associated with failure of antineoplastic treatment. This multidrug transporter is currently an important target for drug design[24] but the recent report that nanomolar concentrations of the antitumor drug ET743 can abrogate the transcriptional activation of the *mdr1* gene by different agents[25] brings about the possibility of pharmacological intervention at the transcriptional level,[26] as has indeed been attempted by use of artificial protein constructs incorporating zinc fingers linked to repressor domains.[27]

Our interest in the natural anti-cancer agent ET743,[28] which is known to bind covalently to the exocyclic amino group of guanine in the minor groove of double-stranded DNA (dsDNA), led us to focus on the *mdr1* proximal promoter (Figure 3) which contains a G-rich regulatory site (also known as GC box) to which either Sp1 (GGGGCGTGG)[29–31] or EGR1 (GCGTGGGCT) can bind specifically, the latter mediating promoter activation by 12-*O*-tetradecanoylphorbol-13-acetate (TPA).[32] This TPA-induced response was later shown to be inhibited in K562 cells by WT1.[33]

It is well known that binding of zinc finger proteins such as EGR1 to dsDNA causes this macromolecule to adopt a distinctive structure that is intermediate between *A* and *B*-type DNA, with a characteristically enlarged major groove.[34] Strikingly, when we compared the DNA minor groove of d(AGCGTGGGCG)·d(CGCCCACGCT) in complex with the three zinc fingers of transcription factor EGR1/Zif268[35] with those of d(TAAAGCTTA)·d(TAAGCTTTA) and d(TAAC GGTTA)·d(TAACCGTTA) in their covalent complexes with ET743 (covalently modified guanine bases underlined),[36] a remarkable structural similarity was revealed that led us to propose that the drug might be selectively recognizing a DNA stretch that is already preorganized for binding upon association with a specific zinc finger-containing protein.[37] This similarity was recently proposed to extend beyond a single ET743-binding site when the feasibility of head-to-tail tandem binding of three ET743 molecules to three adjacent optimal-binding sites (TGG CGG CGG) was demonstrated using unrestrained molecular dynamics simulations.[38] This method makes use of recent advances in force-field parameters, simulation algorithms, and computer power to derive reliable trajectories in atomic detail for macromolecular systems, including protein–DNA complexes,[39] exclusively in terms of intermolecular forces and motions. Analyses of these trajectories can provide independent accounts of experimentally observed behavior, including DNA bending[36] and other sequence-dependent structural effects in DNA duplexes, even when starting from incorrectly modeled structures.[40]

Although the structure of the EGR1 DBD has been reported in complex with a number of optimal and mutated DNA target sites,[5] the only 3D

structures available for Sp1 are those of individual zinc fingers 2 and 3 in solution.[41] Sequence alignment of mammalian Sp1, WT1, and other zinc finger-containing transcription factors such as EGR1 reveals a tandem array of three Cys2-His2 zinc finger domains in their C-terminal regions (Figure 1). Structurally, a Cys2-His2 zinc finger is a ββα motif stabilized by chelation of a zinc ion between the thiolate groups of a pair of cysteine residues from the β-sheet and the imidazole rings of a pair of histidine residues from the α-helix.[42] The α-helical portion of each finger is capable of fitting in the major groove of the DNA in such a way that binding of successive fingers, which are connected by Krüppel-type linkers of suitable composition and length, causes the protein to wrap around the DNA. Base recognition capitalizes on three key residues in this α-helical "reading head", and each zinc finger domain typically binds three successive base-pairs of DNA sequence. For a typical zinc finger, it is well established that the majority of contacts are between the side-chains of helical residues −1 (the residue immediately preceding the α-helix), 3 and 6 with the G-rich strand (primary strand), whose 5′ → 3′ direction runs antiparallel to the N → C direction of the bound protein. In addition, when residue 2 of the helix is aspartic acid, an extra contact can be established with the exocyclic amino group of either adenine or cytosine in the complementary C-rich strand, effectively extending the length of the binding site one nucleotide further outside the triplet. This fact can account for the approximately fivefold increase in affinity for EGR1 when binding to 5′-GAGGGGGAG(T) compared to 5′-GAGGGGGAG(C), as a consequence of the favorable interaction between the amino group in the major groove of A10′ in the former DNA sequence and the side-chain of Asp2 in the first zinc finger of the protein.[20] It is then a set of 1:1 interactions between residues in four positions (−1, 2, 3 and 6) of each zinc finger α-helix and the functional groups present on the edges of the DNA bases in the major groove that largely dictate the binding specificity. This particular arrangement supports the notion that a "recognition code" exists that correlates specific residues in the recognition helix of the protein with particular bases in the DNA site.[4] The code, however, is known to tolerate some level of degeneracy[43–45] and to be somewhat dependent on both intradomain and finger-to-finger context.[4,42,45] In fact, it has been proposed that the coordinated interaction of successive zinc fingers imposed by their modular arrangement is better described in terms of overlapping 4 bp subsites.[46]

Because each zinc finger domain typically binds 3 bp of DNA sequence, a complete recognition alphabet would require the characterization of four[3] domains. For the 16 zinc finger domains representing the 5′-GNN-3′ subset of this 64 member recognition code (where N is any base) a systematic approach has consisted of phage display selection, followed by refinement by site-directed mutagenesis, and rigorous characterization.[45,47] Besides demonstrating that the identity of the residues at the three helical positions −1, 3, and 6 of a zinc finger domain are typically insufficient to describe in detail the specificity of the domain, this and other phage display libraries (see below) have shown impressive amino acid conservation for recognition of the same nucleotide in different targets and a variable amount of cross-reactivity. For example, Arg6 was always used for exquisite 5′-G subsite recognition and Asp2 was coselected with Arg-1 in all proteins for which the target subsite was GNG. Perhaps more remarkable for our purposes is the fact that these selection experiments showed that proteins selected or designed to bind to one particular sequence bind equally well to other sequences, whereas the opposite is not necessarily true.

Much of the effort devoted to decoding the basis of protein–DNA recognition by zinc fingers has been aimed at increasing the specificity of the interaction.[48,49] This enterprise, on the whole, has been extremely successful, as exemplified by the wilful and specific regulation of endogenous genes with designed transcription factors.[49,50] But in the case of Sp1 the same protein is being used in nature to bind to different DNA sequences with similar affinities, and we were interested in understanding the molecular basis for this apparently lax preference.

To this end, we attempted to rationalize the available experimental evidence regarding Sp1 binding to DNA. For completeness and assessment, we built a molecular model of the whole Sp1 DBD and simulated the dynamic behavior of the complexes of this DBD with their respective DNA target sites within the proximal *mdr1* and *wt1* promoters. In addition, as a means of validating the dynamics protocol, the complex between the DBD of EGR1 and its DNA target site within the proximal *mdr1* promoter was simulated under the same conditions. To facilitate comparison between the wealth of experimental data and the results of our simulations, for each specific item in the next section we comment on our findings following a brief summary of the results reported in the literature.

Results and Discussion

After the equilibration period, the progression of the root-mean-square deviations (rmsd) of the coordinates of the solute atoms, including the zinc ions, with respect to the average structures (Figure 4) showed a remarkably stable behavior (Figure 5). The zinc ions remained firmly coordinated and the zinc finger architecture was preserved for the whole length of the simulation. The absence of drifting to higher rmsd values is indicative of adequate sampling during the data collection period and suggests that the simulations were

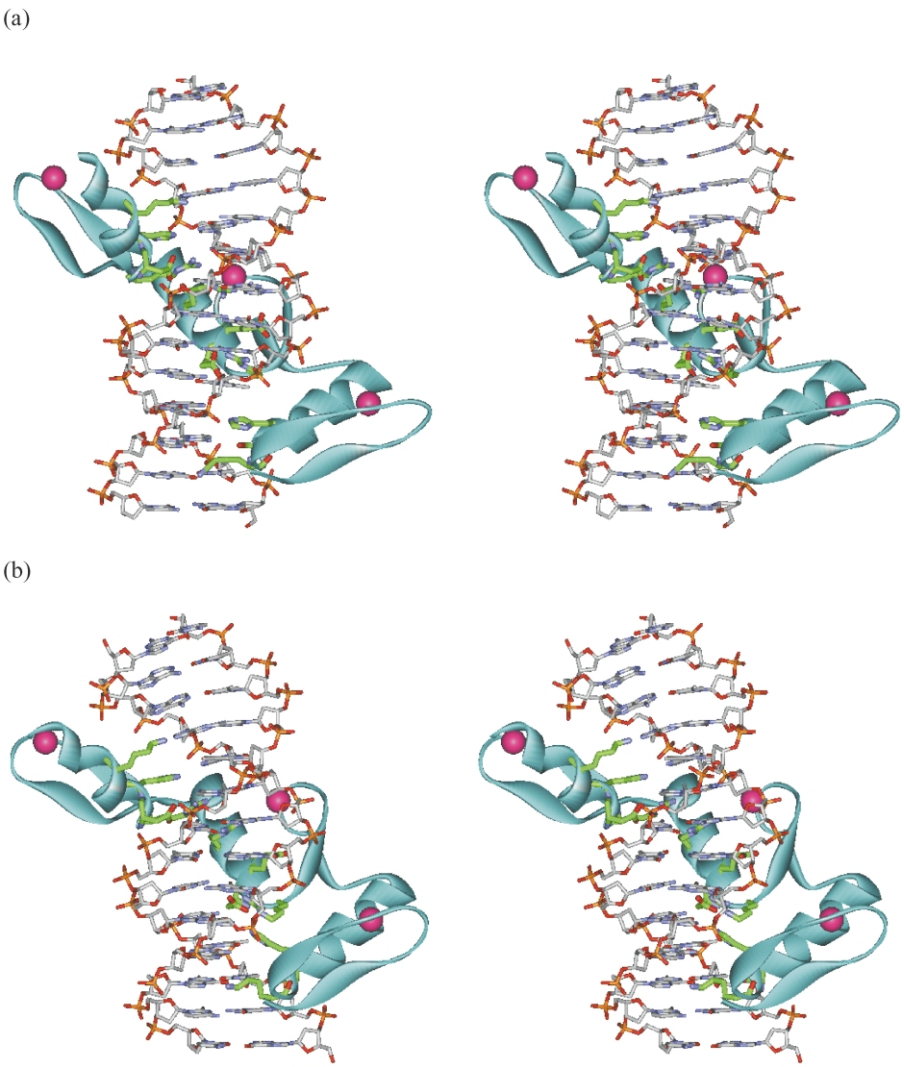

Figure 4. Stereoviews of the energy-refined average structures from the 1000–3000 ps period of the molecular dynamics simulations in aqueous solution of the complexes of Sp1 DBD with their respective DNA-binding sites in the promoter regions of the *mdr1* (top) and *wt1* (bottom) human genes. Zinc ions are shown as purple spheres. For clarity, only the side-chains of residues involved in sequence recognition are shown, with carbon atoms colored in green (Sp1-f1: Lys550, Ser552, and His553; Sp1-f2: Arg580, Asp582, Glu583, and Arg586; Sp1-f3: Arg608, Asp610, His611, and Lys614).

long enough to capture the internal dynamics of the protein–DNA complexes in the three cases studied. The relatively small rms differences calculated for the EGR1–DNA complex with respect to the initial X-ray-based structure (below 2.0 Å) support the validity of the molecular dynamics protocol.

The models obtained for the two Sp1–DNA complexes are consistent with the information about DNA recognition that has been gained from a variety of experimental approaches, including interference of Sp1 binding through methylation of guanine bases,[51] protection of N7 and N3 chemical modification of purines by dimethyl sulfate,[51,52] interference of phosphate ethylation by *N*-nitroso-*N*-ethylurea,[53] base replacements throughout the GC box,[54–56] fragment deletions in the protein,[51,53,56] and amino acid substitutions at critical residues in both the loops and the recognition helices.[57–59] As a result of these studies, the relative importance for specificity and binding affinity of each finger (named Sp1-f1, Sp1-f2 and Sp1-f3), individual protein residues, and DNA base composition (triplet subsites and flanking sequences) has been assessed and is summarized below.

Contacts with the DNA phosphate backbone

Summary of experimental results

Available zinc finger–DNA complexes reveal only three conserved phosphate contacts with the primary (G-rich) strand in the majority of the structures; namely, (i) His at position 7 in the recog-

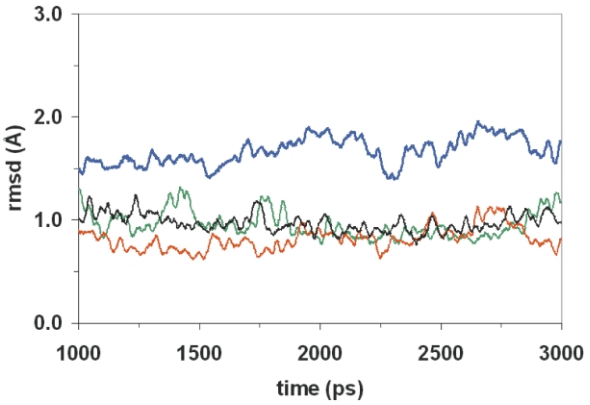

Figure 5. Evolution of the root-mean-square deviations (rmsd) of the coordinates of the solute atoms (including zinc ions) with respect to the calculated average structures (red: EGR1/*mdr1*, black: Sp1/*mdr1*, green: Sp1/*wt1*). The thick line (blue) represents the progression of the calculated rms differences between snapshots taken from the molecular dynamics simulation of the EGR1–DNA complex and the initial X-ray based coordinates. Each line is made up of 1000 individual points.

nition α-helix, (ii) Lys or Arg at position 1 of the second β strand, and (iii) Lys or Arg at the fourth position of the linker joining two consecutive zinc fingers (consensus sequence TGEKP).[42] Additional contacts with phosphate groups from the complementary strand have been reported and have been observed in our simulations (mostly involving amino acid residues at positions +1 and +5 of the α-helix) but these will not be discussed because of their larger fluctuations.

Simulation results

For EGR1 binding to the *mdr1* promoter, our simulation shows that both His3 in zinc fingers 1 and 2 make direct contact with the phosphate groups linking G5 and G6, and C2 and G3, respectively (Table 1). His3 in EGR1-f3 could presumably bind in a similar fashion to the phosphate group joining DNA positions −1 and −2 but direct observation of this interaction is precluded by the limited length of our simulated oligonucleotide (as is the case in all X-ray crystal structures of EGR1/Zif268–DNA complexes). For Sp1 binding to both the *mdr1* and *wt1* promoters, His3 in Sp1-f2 (His583) and Sp1-f3 (His615) maintain this contact during the whole length of the molecular dynamics simulations (Table 1). N^{δ} of His583 hydrogen bonds to the phosphate linking G2 and G3 in the *mdr1* promoter and A2 and G3 in the *wt1* promoter. N^{δ} of His615 does the same with the phosphate groups linking C1 and C2 (*mdr1*), and A1 and A2 (*wt1*). His3 in Sp1-f1, on the other hand, is not close to the backbone but a stable hydrogen bond is formed between the hydroxyl group of the tyrosine residue found at position −3 of the α-helix (Tyr548) and the phosphate linking A5 and G6, and C5 and G6 in the *wt1* and *mdr1* promoters, respectively (Table 1). This contact can be observed in the complexes of DNA with other zinc fingers, such as GLI-f3, GLI-f5, TFIIIA-f1 and TFVIA-f1.

The amino acid residue at position 1 in the second β-strand of each zinc finger is always Arg in EGR1 (R114, R142, and R170) and Lys in Sp1 (K546, K576, and K604) but crystallographic evidence of a direct contact is available only for R114. In agreement with this, our simulation with EGR1 DBD shows a stable hydrogen bonding interaction only between R114 and the phosphate group linking G6 and G7, even though the guanidinium groups of R142 and R170 are within 5 Å of backbone phosphate groups (Table 1). With respect to the protonated amino group of the Sp1 lysine residues, competition with water molecules and the more flexible nature of the side-chain of this residue appear to prevent stable interactions with the backbone and only occasional encounters are observed (data not shown) although the long-range electrostatic interaction can contribute to the binding affinity.

There is no complete identity in the residues making up the linker region that connects adjacent zinc fingers in EGR1 and Sp1 (Figure 1) but in all cases a basic residue occupies position 4: TG(E/Q)(K/R)(K/P). Our simulations show stable direct contacts only for K133 in the first EGR1 linker (with the phosphate group joining G3 and T4), in agreement with the X-ray data, and for R565 in the first Sp1 linker (with the phosphate group joining G3 and G4 in both the *mdr1* and *wt1* promoters). In addition, for K595 hydrogen bonding to the phosphate groups linking G1 and A1 (*wt1*) or G1 and C1 (*mdr1*) is interspersed with complete solvation of the amino group. Both interactions involving R565 and K595, as well as a lack of interaction for K533 and K534, are supported by results from ethylation interference experiments using wild-type and mutant Sp1 peptides.[53]

Besides the aforementioned "conserved" contacts, our simulations for the Sp1 DBD suggest two additional contacts with phosphate groups from the primary strand, involving K535 and W560. K535 precedes zinc finger 1 in the primary sequence and is located at a position equivalent to R565 and K595 relative to the coordinating cysteine residues. This residue, which binds to the phosphate group linking positions 7 and 8, and whose dynamic behavior is similar to that of R565, has been suggested to protect phosphate ethylation.[53] In Sp1-f1, the indole nitrogen atom of W560 (found at position +10 in the α-helix) establishes a permanent (*mdr1*) or slightly fluctuating (*wt1*) hydrogen bond with the phosphate group linking positions 4 and 5 in the G-rich strand. The tryptophan ring is stabilized by stacking interaction with H559 and could provide a docking platform for E2F and other Sp1 binding partners.[60]

Table 1. Intermolecular hydrogen bonding donor–acceptor distances (Å) and angles (degrees) observed in the modeled complexes studied between protein residues and DNA phosphate backbone atoms

Sp1						EGR1		
GAA**GAGGAGGAG**CC			GCC**GGGGCGTGG**GC			**GCGTGGGCT**G		
	Mean distance ± s.d.	Angle values ± s.d.		Mean distance ± s.d.	Angle values ± s.d.		Mean distance ± s.d.	Angle values ± s.d.
Conserved contacts								
His615 ND1 A(−1) O2P	3.2 ± 0.5	155.1 ± 14.0	His615 ND1 C(−1) O2P	3.1 ± 0.5	158.1 ± 12.6			
His587 ND1 G3 O2P	3.3 ± 0.7	134.0 ± 10.1	His587 ND1 G3 O2P	2.9 ± 0.3	159.3 ± 10.8	His153 ND1 G3 O2P	2.8 ± 0.1	161.1 ± 9.8
Tyr548 OH G6 O1P	3.0 ± 0.5	151.5 ± 17.0	Tyr548 OH G6 O1P	2.9 ± 0.5	157.1 ± 14.6	His125 ND1 G6 O2P	3.0 ± 0.3	159.9 ± 11.7
Lys604 NZ A(−1) O1P	3.3 ± 1.0	ND	Lys604 NZ C(−1) O1P	3.6 ± 1.2	ND	Arg114 NH2 G6 O3′	3.2 ± 0.3	145.7 ± 14.6
Lys535 NZ A8 O1P	2.9 ± 0.2	ND	Lys535 NZ G8 O1P	2.9 ± 0.2	ND	Lys133 NZ T4 O1P	3.4 ± 0.9	ND
Arg565 NH1 G4 O1P	3.1 ± 0.3	140.2 ± 14.3	Arg565 NH1 G4 O1P	3.1 ± 0.4	144.6 ± 19.8	Lys161 NZ C2 O1P	5.4 ± 1.2	ND
Lys595 NZ G1O1P[a]	3.6 ± 1.4	ND	Lys595 NZ G1 O1P[a]	3.7 ± 1.1	ND			
Lys595 NZ G1 O1P[b]	5.3 ± 1.2	ND	Lys595 NZ G1 O1P[b]	6.2 ± 1.5	ND			
Non-conserved contacts								
Trp560 NE1 A5 O1P	3.9 ± 1.0	136.2 ± 25.0	Trp560 NE1 C5 O1P	3.1 ± 0.3	138.7 ± 21.9			
Arg555 NH1 G10′ O1P	3.0 ± 0.5	149.7 ± 17.4	Arg555 NH1 C10′ O1P	3.1 ± 0.4	140.1 ± 25.1	Arg155 C7′ NH2	3.4 ± 0.8	157.8 ± 12.0
Ser581 OG T8′ O1P	3.0 ± 1.6	152.3 ± 24.8	Gln585 NE2 C8′ O1P	3.8 ± 0.4	134.3 ± 19.1	Lys179 NZ A4′ O2P	3.6 ± 0.9	ND
Ser581 OG T8′ O1P	3.4 ± 0.9	143.6 ± 30.0	Se-609 OG C5′ O1P	4.0 ± 1.0	141.3 ± 22.1			

Conserved and non-conserved contacts refer to contacts observed in the majority of zinc finger–DNA complexes studied to date. ND, not determined because of alternation among the ammonium hydrogen atoms.

[a] Mean distance during the first nanosecond of simulation.

[b] Mean distance during the rest of the simulation.

Invariant nucleotides in DNA sites recognized by Sp1 DBD

Summary of experimental findings

Box recognition by Sp1 in the simian virus 40 (SV40) early promoter region appeared to be dominated by runs of five guanine bases, G2, G3, G4, G6, and G5′, on both strands in the sequence 5′-GGGCG. Of these guanine bases, G5′ is the only one that is not protected by bound Sp1 from attack by dimethyl sulfate, which reinforces the idea that this base is not involved in close contacts with the protein.[51–53] G1, G7, G8 (in cases where it is not A), and G9, on the other hand, were shown to be contacted by Sp1 only weakly.[51] These findings suggested that the two C-terminal finger domains (Sp1-f2 and Sp1-f3) contribute more strongly to the binding specificity than Sp1-f1, and are consistent with the observation that the GGGCG motif at the 5′-end of the GC box is better conserved than the 3′-portion in Sp1 binding sequences.[1,6] In fact, G is invariably present at position 3 of triplet 1 (G3), and positions 1 and 3 within triplet 2 (G4 and G6) in all Sp1-binding sites reported to date, whereas G1 and G7 can be replaced with T, G2 with A, and G8 and G9 with A, depending on the target sequence that is considered (Figure 2).

Simulations results

In our molecular dynamics trajectories, the only invariant nucleotides (G3, G4, and G6) are systematically recognized by arginine residues and the hydrogen bonding interactions they establish are maintained stably throughout the whole length of the simulations (Table 2).

Binding of Sp1 and EGR1 to adenine-containing sites that diverge from the standard consensus motifs

Summary of experimental findings

It is clearly apparent from the collection of well-characterized Sp1-binding sites (Figure 2) that recognition by this transcription factor is less stringent than that suggested by the commonly accepted GGGCG motif. The most dramatic example is provided by the identification of a single high-affinity Sp1 site as the major enhancer element in the *wt1* gene (Figure 3).[18] This site was shown, rather surprisingly, to consist of a $(CTC)_3$ repeat (= GAG GAG GAG) that nevertheless bound the same form of Sp1 as the consensus GC box and it did so with equally high affinity. It is therefore remarkable that the $(GAG)_3$ repeat was not selected as a target for recombinant Sp1 in a detection assay that used a pool of oligonucleotides with random bases at 12 positions.[55] Similar CTC/GAG repeats with high affinity for Sp1 have been found in the TATA-less promoters of other genes that are growth-related and tissue-restricted in their expression, such as *egr1*, *c-myc*, *c-myb*, *pdgf-A* chain, *egfr*, and *vav*.[18] As a matter of fact, the occasional GAG sequence is indeed present in some of the triplet subsites early reported to be recognized by Sp1 (Figure 2), and some of the pioneering work did actually recognize Sp1 binding to 5′-GGG GAG GGG(C) as only threefold weaker than binding to 5′-GGG GCG GGG(C).[54] This interchangeability of C with A in some triplets has been demonstrated for other zinc finger proteins, such as EGR1, which binds to GAG TGG GAG and GAG GGG GAG with equally high affinity as to GCG (T/G)GG GCG.[20] Consequently, any model of Sp1 binding to DNA must account for the fact that the GAG triplet can be present in all or some of the recognition subsites. Therefore, the usual searches for the GGGCG Sp1 binding motif in promoter regions will be missing out other likely sequences such as AGGAG (or their complementary sequences in the opposite strand).

Simulation results

The analyses of the trajectories show that His3 from Sp1-f3 can indeed bind equally well to the purine N7 of either A2 (*wt1*) or G2 (*mdr1*), as expected (Table 2). The behavior of His3 from Sp1-f1 is somewhat different, however, as it establishes direct hydrogen bonds with the N7 of either G8 (*mdr1*), also in the middle of the triplet, or G7 (*wt1*), because in this latter case A8 can be recognized by Lys-1. On the other hand, the carboxylate group of Glu3 from Sp1-f2, while discriminating strongly against G on electrostatic grounds, can actually accept A5 in the *wt1* promoter through the intervention of a water molecule, as discussed below. In the *mdr1* promoter-binding site, position 5 is occupied by a cytosine base for which no direct interaction is observed, even though the equivalent Glu3 from EGR1-f3 does get engaged in a hydrogen bond with the amino group of C2 (Table 2).

Rationale for the binding specificity of each Sp1 DBD zinc finger

Given the high degree of homology between individual zinc fingers of the archetypical EGR1/Zif268 transcription factor and those of Sp1 (Figure 1), we assumed that the same framework and canonical docking pattern found in EGR1–DNA complexes could be used to construct our models of Sp1–DNA interaction (see Methodology), in agreement with early suggestions.[56,61] This assumption (Figure 6) is supported by an NMR spectroscopy study of Sp1-f2 and Sp1-f3, which revealed 3D structures and key amino acid residues in the α-helices similar to those of fingers 1 and 2 of EGR1/Zif268.[41] Also, the presence of Asp2 in both Sp1-f2 (residue 582) and Sp1-f3 (residue 610) suggests that each of these two residues can bind to an amino group in the major groove at

Table 2. Intermolecular hydrogen bonding donor–acceptor distances (Å) and angles (degrees) observed in the modeled complexes studied between key zinc finger residues and DNA atoms in the floor of the major groove

	Sp1						EGR1			
	GAGGAGGAGC			**GGGGCGTGGG**			**GCGTGGGCT**G			
		Mean distance ± s.d.	Mean angle ± s.d.		Mean distance ± s.d	Mean angle ± s.d.			Mean distance ± s.d.	Mean angle ± s.d.
Finger3 **RSDHLSK**	Lys6 NZ G1 O6	3.1 ± 0.2	ND	Lys6 NZ G1 O6	3.1 ± 0.4	ND	Finger3 **RSDERKR**	Arg6 NH1 G1 O6	3.3 ± 0.6	136.6 ± 16.5
	Lys6 NZ G1 N7	3.1 ± 0.3	ND	Lys6 NZ G1 N7	3.1 ± 0.4	ND		Arg-1 NH1 G3 N7	2.8 ± 0.1	157.8 ± 10.5
	His3 NE2 A2 N7	2.9 ± 0.1	143.4 ± 15.4	His3 NE2 G2 N7	3.0 ± 0.2	147.2 ± 16.8		Arg-1 NH2 G3 O6	2.9 ± 0.1	159.4 ± 9.7
	Arg-1 NH1 G3 N7	3.0 ± 0.2	145.4 ± 13.4	Arg-1 NH1 G3 N7	3.1 ± 0.2	149.6 ± 14.52		Glu3 OE1 C2 N4	2.8 ± 0.1	164.1 ± 8.3
	Arg-1 NH2 G3 O6	3.0 ± 0.2	155.6 ± 11.7	Arg-1 NH2 G3 O6	2.8 ± 0.1	157.3 ± 13.31		Asp2 OD2 A4′ N6	3.1 ± 0.3	152.9 ± 11.4
	Asp2 OD2 C4′ N4	2.8 ± 0.2	139.5 ± 24.9	Asp2 OD2 C4′ N4	3.5 ± 0.6	158.0 ± 14.2				
Finger2 **RSDELQR**	Arg6 NH1 G4 O6	2.9 ± 0.1	144.1 ± 18.4	Arg6 NH1 G4 N7	3.0 ± 0.2	137.4 ± 21.5	Finger2 **RSDHLTT**	His3 NE1 G5 O6	2.9 ± 0.1	160.4 ± 9.9
	Arg6 NH2 G4 N7	3.0 ± 0.1	154.4 ± 14.4	Arg-1 NH1 G6 N7	3.1 ± 0.2	152.3 ± 12.7		Arg-1 NH1 G6 N7	3.0 ± 0.1	145.7 ± 12.8
	Arg-1 NH1 G6 N7	2.9 ± 0.2	154.2 ± 12.9	Arg-1 NH2 G6 O6	2.8 ± 0.1	158.2 ± 13.4		Arg-1 NH2 G6 O6	2.8 ± 0.2	158.0 ± 10.9
	Arg-1 NH2 G6 O6	2.8 ± 0.1	155.8 ± 12.8	Asp2 OD2 A7′ N6	3.4 ± 0.3	150.6 ± 14.0		Asp2 OD2 C7′ O4	2.8 ± 0.1	163.6 ± 8.5
	Asp2 OD2 C7′ N4	2.9 ± 0.2	157.6 ± 15.6							
Finger1 **KTSHLRA**	His3 NE2 G7 O6	3.1 ± 0.3	136.4 ± 15.4	His3 NE2 G8 O6	2.8 ± 0.7	151.3 ± 12.3	Finger1 **RSDELTR**	Arg6NH1 G7 O6	2.9 ± 0.2	146.1 ± 16.1
	Lys-1 NZ N7 G9	3.2 ± 0.3	ND	Lys-1 NZ G9 N7	3.2 ± 0.6	ND		Arg6NH2 G7 N7	2.9 ± 0.1	156.1 ± 14.1
	Lys-1 NZ O6 G9	3.2 ± 0.5	ND	Lys-1 NZ G10 O6	3.0 ± 0.4	ND		Arg-1 NH1 T9 O4	3.2 ± 0.5	136.9 ± 22.1
	Lys-1 NZ N7 A8[a]	3.2 ± 0.4	ND	Ser2 OG C9′ N4[b]	3.3 ± 0.4	141.8 ± 22.0		Asp2 OD2 C10′ O4	3.0 ± 0.3	153.4 ± 17.8
	Ser2 OG C9′ N4	3.0 ± 0.2	154.3 ± 14.2							

Mean distance from 1200 ps to 2000 ps.
[a] This contact was achieved during the last nanosecond of the simulation.
[b] ND, Not determined because of alternation among the ammonium hydrogens.

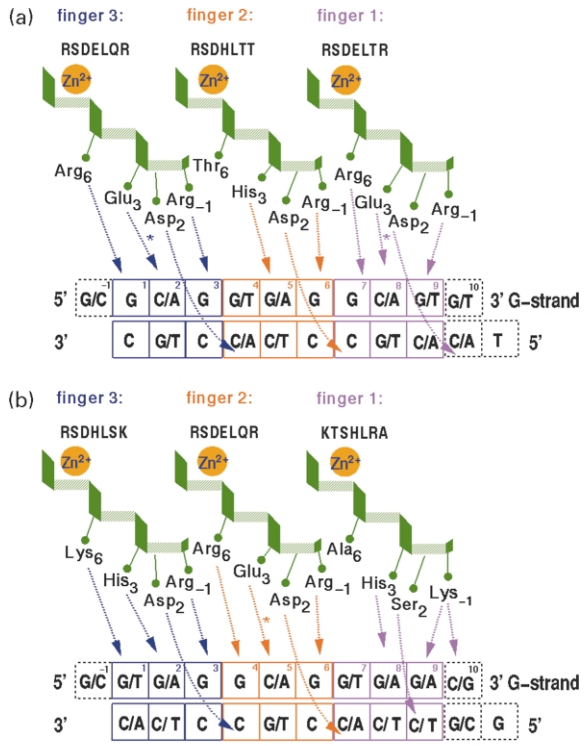

Figure 6. A diagram of modular recognition between DNA triplet subsites containing allowed bases and the three zinc fingers of (a) EGR1 and (b) Sp1 (scheme adapted from Jamieson *et al*.[44] and Isalan *et al*.[46]). Broken arrows indicate direct contacts, whereas the asterisks denote discrimination through electrostatic repulsion. Broken boxes enclose bases outside the minimal nonamer-binding site.

the positions of the first base in the third and second triplets, respectively, in the secondary strand, be it a C or an A (complementary to G or T in the G-rich strand),[59] consistent with findings reported for Zif268.[46] This is particularly interesting, because the first base-pair in the third DNA triplet recognized by Sp1 (usually G, as in the case of the *dhfr* and *wt1* promoters, but also T, as in the *mdr1* promoter) cannot be distinguished uniquely by the residue at position −1 of the α-helix (Ala556) of Sp1-f1, the only Sp1 finger whose 3D structure has not been reported. Therefore, we would ascribe a rather permissive role to Ala6 in Sp1-f1, in as much as it can allow recognition of either T or G in position 7 of the box. In contrast, the presence of Arg6 in Sp1-f2 precludes A and strongly favors G in position 4, which indeed is a conserved G in the G-rich strand in all high-affinity Sp1-binding sites. Additional experimental support for this contention comes from the fact that no clone enrichment was achieved when this base was replaced with A, C or T in phage display selection experiments.[44] Given these differences between fingers 2 and 3, on the one hand, and finger 1, on the other hand, this latter finger will be considered separately from the other two.

Zinc fingers 2 and 3 of Sp1

Summary of experimental findings. Mutation of the helical domain of Sp1-f3 to SSHLIQ significantly impairs binding of the peptide to the target site 5′-GGGGCGGGG-3′.[57] Likewise, replacement of Arg-1 with Ala in fingers 2 and 3 results in a great loss of affinity for the *dhfr* promoter (160-fold and 1100-fold, respectively),[59] as would be expected from the loss of interaction between the guanidinium groups of these arginine residues and the electronegative N7 and O6 atoms of guanine bases G6 and G3, respectively (Table 2). Indeed, positions 3 of triplet 1, and positions 1 and 3 within triplet 2 (i.e. positions 4 and 6 within the box, Figure 6) are conserved as G in all Sp1-binding sites (Figure 2), and our simulations, as commented above, provide additional evidence suggesting that these bases are being recognized by Arg residues in a manner analogous with that in the Zif268–DNA complex (Table 2).[35] On the other hand, the presence of Lys, and not Arg, at position 6 of Sp1-f3 (residue 614) can mean only that G is not absolutely required in this position because Arg6 is used systematically for exquisite 5′-G recognition.[4,42]

With regard to the middle position in these DNA subsites, it is interesting to note that in phage display experiments the sequences **RS*D*HLTR** and **RK*D*SLVR**, selected for binding to GGG and GTG, respectively, were shown to bind GAG with similar affinities, whereas **RS*D*NLRR** and **RS*D*NLVR**, both selected to bind GAG, bound GTG or GGG much more weakly.[45] Likewise, **RS*DD*LVR** and **RL*D*TLGR**, both selected to bind GCG, happened to recognize GAG with even greater affinity, whereas those selected to bind GAG were more discriminatory against GCG. In view of the relative variety of sequences to which Sp1 binds with high affinity (Figure 2), it is very interesting that His3, always selected to recognize the central G, is shown not to discriminate against A, as opposed to Lys. Thus, the sequence **RS*D*KLVR**, highly specific for GGG, would not be suitable for recognition of GAG.

Simulation results. Arg-1 (residue 580) and Arg6 (residue 586) in Sp1-f2, and Arg-1 (residue 608) in Sp1-f3 recognize, respectively, the bold-face typed bases of the central **G**(C/A)**G** subsite and the 3′ **G** in the first triplet subsite. These G3, G4 and G6 are precisely the invariant nucleotides discussed above. The side-chain of Lys6 in Sp1-f3 effectively appears to be long enough to establish a direct interaction with a guanine base in the 5′-base of the first triplet (as in both **G**GG and **G**AG) (Table 2). Given the equivalent location in the floor of the DNA major groove of guanine N7 and thymine O4, a thymine base at this position (i.e. **T**GG) could be recognized equally by this lysine. Also in agreement with the previous reasoning, our molecular dynamics simulations with the *mdr1* and *wt1* promoters show a similar hydrogen bonding

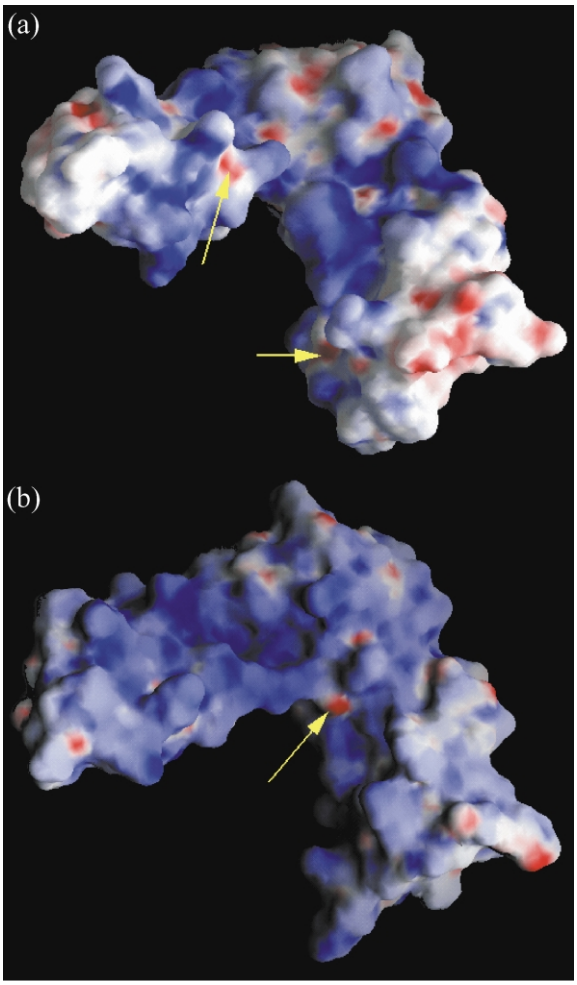

Figure 7. Molecular surface representation of the zinc-finger domains of EGR1 (top) and Sp1 (bottom), color-coded according to electrostatic potential (blue, most positive; red, most negative, with intermediate values ramping smoothly) as obtained by using program GRASP (http://honiglab.cpmc.columbia.edu/grasp/) (A. Nicholls & B. Honig, Columbia University, unpublished) and the charges used in the force-field. The yellow arrows highlight the negative regions surrounding the side-chains of the glutamate residues that are strongly discriminant against guanine in the respective DNA target subsite. Note that the location of this residue differs between EGR1 (fingers 1 and 3) and Sp1 (finger 2), which is likely to be of importance when these transcription factors compete for binding to overlapping sites.

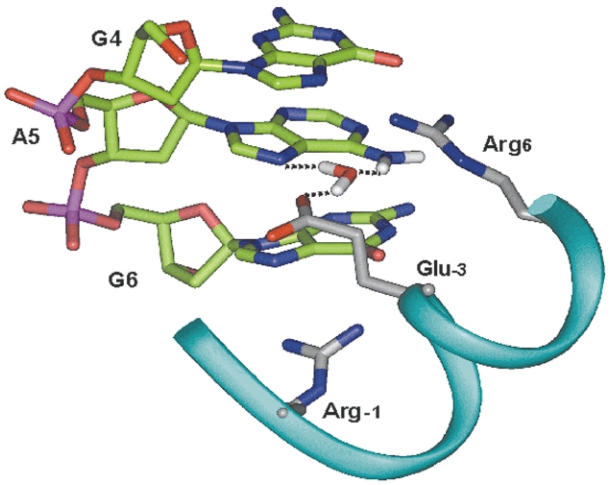

Figure 8. Detail of the structure of the complex between Sp1 DBD and its DNA target site on the *wt1* promoter showing the proposed water-mediated interaction between Glu3 in the second zinc finger and A5 in the middle position of the central triplet. For clarity, only hydrogen atoms involved in this particular interaction are shown. This water molecule, which acts both as a hydrogen bond donor and acceptor, shows a long residence time (>80% of simulation time).

pattern between the N^{ϵ} of His3 and N7 of either G2 or A2 (Table 2).

Since we now know that the α-helices of Sp1-f2 (**RS*D*ELQR**) and Sp1-f3 (**RS*D*HLSK**) can recognize, respectively, GAG/GCG and GAG/GGG/TGG with similar affinities,[43,47] we would expect these helices to resemble more closely **RS*D*HLTR** than either **RS*D*NLRR** or **RS*D*KLVR**, and this is indeed the case. In other words, if we were to choose a peptide sequence that could recognize equally well the triplets GGG and GAG, the sequence **RS*D*HLTR** should be preferred over **RS*D*NLRR**, whereas to bind with high specificity to GAG, the latter peptide would be chosen.

On the other hand, a glutamic acid residue at position 3 of the recognition helix is known to confer relatively low specificity[62] with just a slight preference for C but strong selection against G.[20,63] In our models, Glu3 (residue 583) of Sp1-f2 packs its methylene β-protons against Phe578 and does not contact the floor of the DNA major groove directly (as seen in the X-ray crystal structures of DNA-bound EGR1-f1 and EGR1-f3). However, molecular electrostatic potential (MEP) calculations clearly show a negative MEP region around the buried carboxylate group of this amino acid residue in the middle of the helix (Figure 7) that would interact repulsively with the negative MEP region surrounding the N7 and O6 region of a putative guanine base at position 5 in the primary strand. The question remains, however, of why an adenine base is accepted at this position, given that this purine has an electronegative N7 in the same position as that in guanine. The answer is provided by our simulation of the complex containing the central GAG, which shows that a water molecule is stabilized in a very definite orientation to mediate the interaction between the carboxylate group of Glu583 and the N7 of A5 (Figure 8).

Taking all of these findings together, the second and third zinc fingers of Sp1 appear to have evolved to achieve the goal of binding with similar affinities to subsites G(A/C)G and either G(G/A)G or (G/T)GG, respectively.

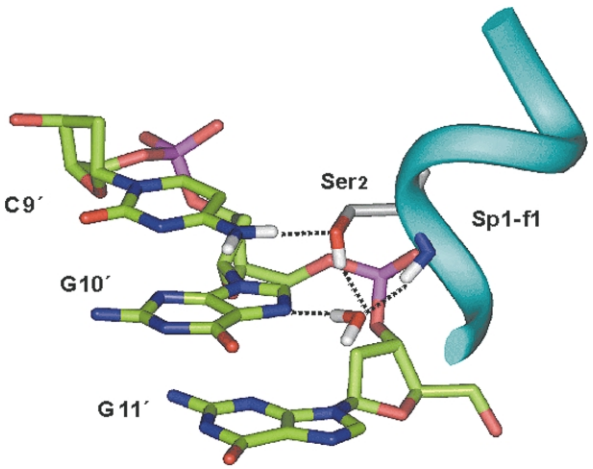

Figure 9. Detail of the structure of the complex between Sp1 DBD and its DNA target site on the *wt1* promoter showing the proposed water-mediated interaction between Ser2 in the first zinc finger and G10′ outside the minimal nonanucleotide-binding site. For clarity, only hydrogen atoms involved in this particular interaction are shown. This water molecule exchanges with neighboring water molecules during the simulation.

Zinc finger 1 of Sp1

It is generally accepted that the sequence selectivity of Sp1-f1 is more relaxed than those of Sp1-f2 and Sp1-f3, and that the role of this finger in the DNA-binding properties of Sp1 is less clear, due to contradictory evidence. Thus, in one experiment it was reported that a peptide containing only Sp1-f2 and Sp1-f3 bound strongly to the GC box of the metallothionein promoter, implying that the third triplet is not crucial for binding.[56] Later, it was shown that deletion of Sp1-f1 resulted in an increase of K_d values from ~3 nM to ~300 nM,[59] suggesting that this finger provides an important contribution to affinity toward the GC box, albeit probably through non-specific interactions. On the other hand, the results of methylation interference analysis reported above suggest that Sp1-f1 does indeed interact with a 5 bp DNA subsite at the 3′-portion of the GC box.[51]

One aspect that is clear, however, is that the composition of the α-helix of Sp1-f1 (**K**TS**H**LR**A**) is rather different from that of helices in fingers 2 and 3. The main differences are (i) the presence of Ala in position 6, which is usually employed to recognize the 5′ base in the triplet subsite, (ii) the replacement of Arg by Lys in position −1, and (iii) the lack of Asp in position 2. We will now deal with each of these issues in light of the experimental evidence and the simulation results.

Ala in position 6 of the Sp1-f1 helix. Although Ala could, in principle, be thought of as not being involved in base recognition, phage display selection experiments have shown at least three instances in which Ala was selected in position 3 of the recognition helix: **D**PG**A**LV**R**, **T**SG**A**LT**T** and **R**KD**A**LT**R** for target sites GTC, GTT and GTG, respectively.[45] This finding implies that the methyl group of Ala, a residue with low recognition ability *per se*, is a suitable functional group to face the methyl group of a thymine base in the major groove, in good accord with earlier experiments.[44] On the other hand, Ala6 of this α-helix might be playing a minor or even superfluous role in Sp1-f1, given the known overlap of neighboring subsites and the possibility of recognition of the base in the complementary strand (position 7′ in the box) by Asp2 in Sp1-f2.

This latter hypothesis is corroborated by our molecular dynamics results showing that Ala6 (residue 556) is too far from the major groove to play any role in specific recognition of the G-rich strand, whereas the carboxylate group of Asp582 is within hydrogen bonding distance from the amino group of either C7′ (*wt1*) or A7′ (*mdr1*) in the complementary strand (Table 2). This interaction can indeed account for the conserved preference for G or T in position 7 in all high-affinity Sp1 sites (Figure 2). The hydrophobic environment created by the side-chain of Ala6, on the other hand, might explain why G7 can be hypermethylated in the presence of bound Sp1 (see below).[51]

Lys in position −1 of the Sp1-f1 helix. The presence of Lys rather than Arg in position −1 of Sp1-f1 is, in a manner similar to that of Lys6 in Sp1-f3, suggestive of less stringency for G in the selection of the last base in the third triplet (both G and A can be found at this position), again in good agreement with sequence preferences (a GGA third triplet is present, for example, in the *HIV-LTR* II and *p21* promoters). Moreover, Lys-1 has been reported to specify 3′ A in a GGA context.[44] In any case, the fact that replacement of this Lys by Ala leads to a much smaller loss in affinity for the GC box of the *dhfr* promoter (~7-fold) than the equivalent substitution in Sp1-f2 and Sp1-f3[59] implies that this interaction is less important for high-affinity binding than the corresponding interactions emanating from Arg-1 in Sp1-f2 and Sp1-f3. In this respect, it has been noted that deletion of individual base-specific interactions at the ends of a zinc finger array, as accomplished by site-directed mutagenesis, has a relatively small effect on complex stability.[4]

In our initial EGR1/Zif268 framework for the Sp1 DBD, Lys-1 in Sp1-f1 appeared to play a dominant role in the recognition of the relatively conserved G9 (it could well be A9, as for example in the *p21* promoter) but during the molecular dynamics trajectories we observed that the amino group of this lysine residue could interact with N7/O6 of G10 (*mdr1*) or N7 of A8 (*wt1*) in the same DNA strand. These fluctuations between neighboring positions are therefore different for TGG or GAG triplet subsites and are in consonance with the reduced methylation interference of G8 and G9 that was observed for a GGG subsite when Lys-1 was changed to alanine.[59]

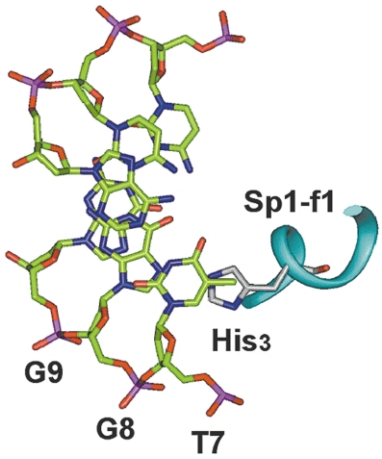

Figure 10. Detail of the structure of the complex between Sp1 DBD and its DNA target site on the *mdr1* promoter showing the proposed stacking interaction between the methyl group of T7 in the third triplet subsite and the imidazole ring of His3 in the first zinc finger. For clarity, all the hydrogen atoms have been omitted.

Lack of Asp in position 2 of the Sp1-f1 helix. Replacement of Ser552 (Ser2 in Sp1-f1) by Ala was shown not to change significantly the dissociation constant of Sp1(530-623) to a canonical GC box[59] but more recent work based on a Ser → Asp mutation has suggested a modest contribution of this position to both binding affinity and base specificity.[64] As discussed below, the hydroxyl group of the side-chain of Ser2 in Sp1-f1 is capable of receiving a hydrogen bond from the amino group of C9′ (present in both promoters) and, through the concourse of a bridging water molecule (Figure 9), of donating a hydrogen bond to the N7 atom of G10′ (as in the *wt1* promoter).

Role of His in position 3 of the Sp1-f1 helix. His in position 3 can recognize either G or A in the middle of the triplet, according to the phage display selection experiments described above,[45] and in agreement with structural studies.[42] Although a mutant containing Ala in place of His at position 3 of this finger has been reported to have practically the same affinity as wild-type Sp1 for a canonical GC box,[59] the authors did not address the issue of specificity, so that a definite role for His3 in base recognition cannot be ruled out. Other experiments have shown that mutation to Ala of similarly highly conserved residues such as Arg6 in finger 3 of EGR1[65] or in finger 2 of ADR1[66] (which is involved in both cases in binding to a guanine base according to the structural studies) can result in only a moderate reduction in complex stability. It is therefore likely that in Sp1-f1 the contribution of this histidine to the energetics of DNA binding may be not so great as that of the corresponding residue in finger 3. A role in sequence recognition, on the other hand, cannot be ruled out and is indirectly supported by the fact that the equivalent position in Sp2-f1 is a leucine (Figure 1) and Sp2 binds to a GT-rich element rather than to the classical GC-boxes to which Sp1 binds.[1]

Our simulations shed some light on this issue as well: His3 is seen to donate a hydrogen bond to N7 of G8 in the *mdr1* promoter (TGGGC subsite) that appears to be stabilized by the concomitant existence of a stacking interaction between the imidazole ring and the methyl group of the preceding T7 (Figure 10). In the case of the *wt1* promoter, on the other hand, His3 is incapable of interacting directly with N7 of A8 (GAG CC subsite) but it does interact with N7 of G7 instead. This is probably a consequence of (i) the larger mobility of its side-chain due to the lack of an equivalent supporting stacking interaction by guanine and (ii) differences in roll angle between TpG and GpA steps. We can assume that a similar behavior would be observed in the case of the subsite that was explored experimentally (GGG CC).[59] In light of these findings, it can be thought that a thymine base at position 5′ of the third subsite will be more favorable for Sp1-f1 binding than a guanine base. This preference is reminiscent of that observed for EGR1-f2, which was shown to select for T more frequently than for G at position 5′ in the middle subsite.[20]

DNA sequence recognition by the three zinc fingers of Sp1 DBD

Taken together, all the available evidence suggests that the DBD of Sp1 consists of a Zif268-like arrangement of three zinc fingers with the capacity to recognize three consecutive DNA triplets through the concourse of the three recognition helices. Preferred sequences making up the first triplet are TGG, GGG and GAG. The third triplet can be TGG, GGG, GAG, GGA, or even GAA, whereas the second is more restrictive, in that only GCG or GAG can be recognized specifically and GGG is excluded strongly (Figure 4). This means that Sp1-f1 can, in principle, bind in a canonical arrangement to the reported high-affinity subsites without the need to invoke any special mechanism beyond the realm of the accepted recognition code. The origin of the lower affinity of this finger relative to fingers 2 and 3 may be accounted for by the presence of Ala in a position frequently occupied by Arg and strongly associated with high-affinity binding.[44] Intriguingly, not all combinations of these subsites appear to be allowed, and some are more favored than others. Thus, GAG GCG TGG is found in the third Sp1-binding site on the HIV-LTR promoter, and the reverse TGG GCG GAG is found in the third and fifth Sp1-binding sites of the SV40 promoter (Figure 2). Despite being regarded as similar high-affinity sites, the former has been shown to bind Sp1 with higher affinity than the latter when nested within identical flanking sequences.[56] A role for His3 in this preference has just been discussed. An inverted GAG GCG GAG(C) Sp1-binding site has been identified only recently in the

proximal promoter of the human gene encoding telomerase reverse transcriptase (hTERT), where it plays a role in transcriptional repression rather than activation.[67] On the other hand, to the best of our knowledge, alternative sequences such as TGG GAG TGG or GCG GAG GCG do not appear to have been identified so far in naturally occurring Sp1-binding sites, and a systematic sequence analysis of this sort has not been reported in the literature. Nonetheless, an elegant "selection and amplification of binding" experiment[20] has shown recently that the (GGG)TGG GCG TGG(C) sequence is selected by Sp1(530-623), to which it binds with 2.3-fold higher affinity than for the more "standard" (GGG)GGG GCG GGG(C).[68]

More difficult to explain is the notable reduction in affinity (~80%) and promoter activity (~65%) that resulted when GAG substituted for the central GCG in the wild-type human *mdr1* promoter,[29] despite the theoretical equivalence of adenine and cytosine in the middle position of the box and the existence of identical sites of reportedly high affinity in other promoters (e.g. HIV-LTR I). Likewise, replacement of the sequence 5′-AGGGCGTGG-3′ in the wild-type Chinese hamster *dhfr* minimum promoter with 5′-AGGGAGTGG-3′ resulted in a threefold decrease in Sp1-binding affinity and ~50% reduction in RNA levels.[69] Similarly, replacement in the same *dhfr* promoter of the TGG triplet recognized by Sp1-f1 by TAG resulted in a tenfold reduction in protein binding.[69] These examples demonstrate that, in practice, base recognition can be limited by differences in the context of the binding triplet and/or the experimental conditions.

If a limited number of combinations do exist, additional factors could be at play, such as sequence-dependent variation in DNA structure (e.g. narrower groove width in A/T regions, intrinsic tendency of CpG and TpG steps to bend into the major groove,[36,38] twist/roll preferences[38] and greater rigidity of ApG and GpA steps compared to CpG and GpC steps[70]). In fact, it is now accepted that the best residues for DNA recognition can depend on the position of a finger in the protein and/or the effect of neighboring fingers so that each zinc finger is not completely independent.[42,44,45,47,63,71] This lack of quantitative modularity is compounded by end effects, as distal hydrogen bonds have been proposed not to be as highly constrained or shielded from the solvent as the more central ones.[65] Support for this hypothesis can be gained from inspection of the results for EGR1-f3 from our molecular dynamics simulation, which shows the largest standard deviation for the Arg6-G1 interaction (Table 2). Given the sequence differences between Sp1-f1 and Sp1-f3, their distinctive affinities might be a reflection of diminished sequence specificity on the part of Sp1-f1 and of differential end effects. This can be especially true in the native protein, as the sequence immediately anteceding the first zinc finger of Sp1, which contains several additional cysteine residues, could act as a proximal accessory region in a way similar to that described for the transcription factor ADR1[72] and have an influence on Sp1-f1 binding specificity.

Extension of the Sp1-binding site beyond the nine nucleotide region

Summary of experimental results

A peptide containing Sp1 residues 533-623, called Sp1(533-623), was demonstrated to contain all of the amino acids that are essential for mimicking the tight binding to cognate DNA sequences of the whole protein.[56] This 93 amino acid zinc finger domain of Sp1 has therefore been considered the "minimal DBD". Primer-extension/mobility-shift assays revealed that for maximal binding affinity there exists a requirement for additional nucleotides beyond the end of the consensus sequence at the 5′-end of the GC-box. Methylation interference analyses carried out with a similar Sp1 fragment, Sp1(530-623), and several mutant peptides,[53] revealed that wild-type Sp1-f1 interacts with G7, G8, and G9 in the G-rich strand at the 3′-portion of a 5′-(A)GGGGCGGGG(CC)-3′ box (underlined guanine bases) and with G10′ and G11′ in the complementary strand. Taken together, these findings point to a possible extension of the length of the third DNA subsite recognized by the N-terminal region of these Sp1 peptides to at least 5 bp. Contacts with G5′, G10′, and G11′ similar to those observed for Sp1(530-623) were deduced for mutant peptides Sp1(R565S), Sp1(R565K), Sp1(K595S), and Sp1(K535G) as well but they were reportedly weaker for the deletion mutant Sp1(537-623), highlighting the importance of the DPGKKKQ 530-536 peptide fragment for binding affinity and sequence recognition. Additional indirect evidence for an extension of the length of the Sp1-binding site from 9 nt to at least 10 nt on the 3′-side is available. Thus, a C → A replacement at position 10 only, as in going from 5′-GGGGCGGAG(C)-3′ in the dihydrofolate reductase (*dhfr*) promoter sites II and IV to 5′-GGGGCGGAG(A)-3′ in site I of the SV40 promoter, was shown to bring about a substantial loss in affinity for Sp1.[56] Similarly, a change from C to T at this position from 5′-TGGGCGGGG(C), as in the HSV I-E3 promoter, to 5′-TGGGCGGGG(T), as in the HSV thymidine kinase promoter, determined a decrease in the affinity of the site from high to medium.[56] These findings in relation to position 10 are reminiscent of those reported for EGR1.[20]

Simulation results

Consistent with this experimental evidence, analysis of the molecular dynamics trajectories shows that the carboxylate group of Asp2 from EGR1-f1 can indeed establish a good hydrogen bond with the amino group of C10′ (Table 2),

Table 3. Width of the DNA minor groove in the complexes of Sp1 DBD with the two oligonucleotides studied

	P5-P28	P6-P27	P7-P26	P8-P25	P9-P24	P10-P23	P11-P22	P12-P21	P13-P20	P14-P19
GAA**GAGGAGGAG**CC (*wt1*)	6.3 ± 1.8	6.8 ± 1.5	6.6 ± 1.1	7.1 ± 1.5	7.3 ± 1.2	7.7 ± 0.9	6.7 ± 1.0	5.9 ± 1.2	5.5 ± 1.1	5.1 ± 1.0
GCC**GGGGCGTGG**GC (*mdr1*)	8.0 ± 1.4	9.9 ± 1.1	7.6 ± 1.0	7.1 ± 1.1	7.6 ± 0.8	6.5 ± 1.0	6.9 ± 0.9	6.5 ± 1.0	7.2 ± 1.0	6.9 ± 1.4

The width of the DNA minor groove was measured as the shortest inter-phosphate distances (Å) across the groove (P–P distance minus 5.8 Å). The P–P distance in standard *B*-DNA is 5.9 Å. Mean values obtained from the last 2 ns of the molecular dynamics simulations are shown, together with their standard deviations. Numbering refers to sequence starting from the 5′ base in the G-rich strand (1-14) followed by the complementary strand in the same direction (15-28).

Table 4. Helical twist parameters for the base-pair steps comprising the central nonanucleotide to which the three fingers of Sp1 DBD bind

	wt1 promoter									
	G1-A2	**A2-G3**	**G3-G4**	**G4-A5**	**A5-G6**	**G6-G7**	**G7-A8**	**A8-G9**	**G9-C10**	**C10-C11**
Twist angle (±s.d.)	33.4 ± 4.2	33.0 ± 3.8	36.3 ± 4.8	32.9 ± 4.1	27.4 ± 5.5	38.5 ± 4.3	33.5 ± 3.9	32.0 ± 3.5	27.7 ± 4.0	32.9 ± 4.9
	mdr1 promoter									
	G1-G2	**G2-G3**	**G3-G4**	**G4-C5**	**C5-G6**	**G6-T7**	**T7-G8**	**G8-G9**	**G9-G10**	**G10-C11**
Twist angle (±s.d.)	32.0 ± 5.8	36.0 ± 3.8	33.9 ± 4.6	25.6 ± 4.5	37.7 ± 3.5	34.9 ± 3.7	30.0 ± 4.9	31.4 ± 4.6	32.9 ± 4.9	31.7 ± 5.7

Mean values (deg.) ± standard deviations were obtained from the last 2 ns of the molecular dynamics simulations. Numbering refers to the central nonanucleotide to which the three fingers bind, plus the 3′ end, in consonance with the main text (see Figure 6).

which accounts for the preference for G at position 10. Since a T can be present at this position (T10), a similar interaction can be foreseen with the amino group of its complementary adenine base (A10′). This reasoning, however, cannot be applied to Sp1-f1, as position 2 in this finger is occupied by a serine residue, whose side-chain hydroxyl group is seen to accept a hydrogen bond from the amino group of C9′ in both the *mdr1* and *wt1* promoters. In the former, G10 can be recognized by Lys-1, which is simultaneously or alternatively binding to G9, whereas in the latter the hydroxyl group of Ser2 is bonded to a water molecule that bridges an interaction with G10′ (Figure 9).

Structural deformations brought about by Sp1 binding

Minor groove width

Hydroxyl radical footprinting of the 21 bp repeats of the SV40 early promoter revealed a regular undulating cleavage pattern consisting of enhancement by Sp1 at the central part of each GC-box region and decreased efficiency between each GC-box.[73] This finding was suggestive of widening of the minor groove and is consistent with the structural information available for zinc fingers bound to oligonucleotides.[42] Monitoring of this width during our simulations as the shortest interphosphate distances across the groove (P–P distances minus 5.8 Å) shows the largest increase in the central region of both promoters, whereas the 3′ region of the *wt1* promoter is the least enlarged (Table 3). To illustrate this point, in the refined averaged structures of our modeled complexes the mean values are 7.4(±0.9) Å (*mdr1*) and 6.5(±0.7) Å (*wt1*), which indeed represent a substantial increase relative to the same distance measured in standard *B*-DNA (5.9 Å).[38]

DNA unwinding

Experiments using a peplomycin–iron complex detected new cleavage sites at the phosphate groups linking either G7 and A8 or (to a lesser extent) G7 and G8 in some of the SV40 GC-boxes (Figure 2).[51] On the other hand, the consistently reported hypermethylation at G7 in the primary strand has been taken by several authors as an indication of Sp1-induced changes in the local conformation of the DNA.[51,52] In fact, the suggested increase in helical twist angle at the G6-G7 base-pair step[51] and a concomitant relative decrease at G7-A8 are observed in our modeled complex of the *wt1* promoter (38.5 ± 3.8 and 33.5 ± 3.6, respectively) but these changes do not appear to make the N7 nitrogen atom of G7 more accessible (or reactive) than in naked DNA. Instead, from inspection of our models, we suggest that it is the presence of both the guanidinium moiety from Arg-1 in finger 2 (Arg580) and the short side-chain of Ala6 in finger 1 (Ala556) that creates a suitable environment for lodging the dimethyl sulfate reagent that alkylates the guanine base.

Sp1-induced DNA unwinding has been reported for a plasmid containing 15 copies of a 5′-GGGGC-GGGG(C)-binding site each separated by 26 bp.[74] It was reasoned that the unwinding was necessary for proper alignment of the protein residues and the DNA interaction sites. Its extent (~18° more per Sp1-binding site than in a control DNA plasmid containing no Sp1-binding site) was found to be in good agreement with that observed in zinc finger protein–DNA cocrystal structures.[35,42] In line with these findings, our own estimations of unwinding for the two modeled Sp1–DNA complexes are 20.1° and 29.8° for the *wt1* and *mdr1* promoters, respectively, and are slightly less than the 32.5° we calculate for the EGR1–DNA complex (which compares well with the 34° measured in the X-ray complex PDB 1aay). These estimates were obtained by adding the average helical twist values corresponding to the 10 bp steps comprising the DBD-binding site (full data set in Table 4) and subtracting 356° corresponding to the sum of twist angles of 10 bp steps of standard *B*-DNA (10 × 35.6°).

DNA bending

Despite this unwinding, no significant bending by Sp1 was detected in circular permutation experiments in which 110 bp restriction fragments with an Sp1-binding site at various locations were used.[74] Nevertheless, the 21 bp repeats of the SV40 early promoter reported above revealed an intrinsic in-phase bent structure that was stabilized by Sp1 binding and a tendency for this region containing six properly spaced Sp1-binding sites to form circular DNA molecules, with each Sp1 molecule contributing ~30° to the overall bending. Additional circular permutation assays carried out by a different group suggested no bending induced by binding of Sp1(537-623) but, intriguingly, some modest bending induced by the longer Sp1(530-623), with the center of the bend located at the center of the GC-box.[53] Early electrophoretic mobility-shift assays, on the other hand, indicated that Sp1 bends DNA upon binding to its recognition sequence, with a bending center at the 5′-side of the GC-box,[75] and the Sp1(533-623) "minimal DNA-binding domain" alone was shown to be capable of producing an asymmetric bending pattern with an even larger bend angle towards the major groove than the full-length protein.[56] On the contrary, a bend center displaced towards the 3′ end of the GC box and ~60° DNA bending into the major groove induced by Sp1(530-623) have been reported.[58] In this latter case, replacement of 5′-GGGGCGGGG(C) by 5′-GGGGCGGAT(A) led to a reduction of ~10° in the bending angle and to a significantly reduced binding affinity, in agreement with earlier results.[56] Given the differences in bending found between Sp1(530-623) and Sp1(537-623), a prominent role in

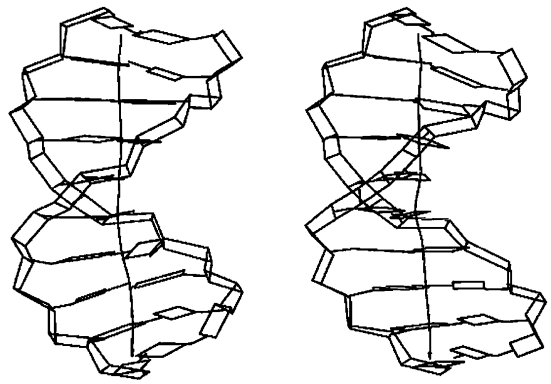

Figure 11. A representation of the DNA double helices as found in the energy-minimized average structures of the modelled EGR1/*mdr1* (left) and Sp1/*mdr1* (right) complexes. Note that the long helical axes, as calculated by CURVES,[97] are virtually straight in both cases.

causing this effect was suggested for Q536.[53] However, the fact that neither the Sp1-induced bending pattern nor the binding affinity of the DBD of this transcription factor was modified upon replacement of Q536 by isoleucine[58] appears to rule out the implication of this residue in the association.

Taken together, these somewhat conflicting findings are indicative of no clear origin for the bending induced by the Sp1-DBD and suggest a significant contribution of Sp1-f1 to specific base recognition and binding affinity, as discussed above, and to the magnitude of the bend angle. Sp1-induced bending can have important biological consequences, as bound Sp1 has been recognized by scanning transmission electron microscopy to associate into tetramers that can subsequently stack in register at a DNA loop juncture. This makes it possible for remote Sp1-binding sites to translocate to the proximal promoter Sp1 sites, which may account for long-range enhancer effects.[76]

In our hands, for the DNA complexes modeled with Sp1(534-623), no significant global curvature is observed for the double helix in the stretch covered by the Sp1 DBD, as assessed by both CURVES and FREEHELIX. In fact, the global helical axis is as straight as that found in the EGR1–DNA complex (Figure 11). The absence of net curvature in these complexes is mostly due to cancellation of localized bends induced by individual fingers over virtually one turn of the helix (the largest positive roll values occur at the most unwound base-pair steps; data not shown). The local helical axes of the base-pairs are writhed around an unbent global axis, resembling *A*-DNA, as discussed.[38] Since our oligonucleotides (14-mers) are much shorter than those used in the experiments, we cannot presently assess the effects of Sp1 binding on the sequences flanking the "GC box" but our results strongly suggest that Sp1-induced bending is more likely to take place at the junctions between GC boxes and intervening sequences than at the GC boxes themselves.

Comparison of the Sp1–DNA and EGR1–DNA complexes obtained from the molecular dynamics simulations

Sp1-f2 and Sp1-f3 each contain an arginine-serine-aspartate tripeptide (RSD motif) that has been well characterized in Zif268-like zinc fingers and is present in all three fingers of EGR1 DBD. Upon complex formation, Arg580 and Arg608 of Sp1 make the expected bidentate contacts to DNA, donating hydrogen bonds to the O6 and N7 atoms of G3 and G6 in the G-rich strands of both *mdr1* and *wt1* promoters. These interactions are similar to those established between Arg146 and Arg174 of the EGR1 DBD with the same DNA bases (Table 2). Besides, both Asp582 and Asp610 of Sp1 buttress the arginine side-chains so as to stabilize their interaction with the electronegative atoms of the guanine bases in the floor of the DNA major groove, as is the case for Asp120, Asp148, and Asp176 in EGR1 DBD (Table 5). At the same time, Asp610 (Sp1-f3) and Asp582 (Sp1-f2) interact, respectively, with the amino groups of C4′ and either A7′ (*mdr1*) or C7′ (*wt1*), each complementary to the 5′ base in the binding site of the preceding finger; the equivalent interactions in EGR1 involve the three fingers (Table 2), effectively extending the recognition site from nine to ten base-pairs (Figure 6). Therefore, although these interactions do not appear to contribute strongly to binding affinity,[44] they do actually play an important role in binding specificity.

The more relaxed sequence selectivity of Sp1-f1 and its lesser contribution to binding affinity and specificity relative to Sp1-f2 and Sp1-f3[59] appear to be a direct consequence of the lack of buttressing interactions in this finger due to the absence of Arg and Asp in the KTS motif in place of the more common RSD motif. As a result, Lys550 recognizes G9 in both promoters but is seen to alternate between G9 and G10 in the *mdr1* promoter (Table 2), whereas the side-chain of Ser552 can bind the amino group of C9′ (directly) or the N7 atom of G10′ (through a water molecule with a short residence time). Incidentally, the presence of Lys-1 in fingers 1 and 3 of WT1 argues in favor of a relative tolerance (either G or A) at the 3′ base in the second and fourth triplet subsites to which the DBD of this transcription factor binds.

The Arg6 interaction with the 5′ guanine base that is associated with high-affinity binding is present in the first and last subsites of the EGR1–DNA complex but only in the central subsite of the Sp1–DNA complexes. Base composition in the middle position of these subsites is governed by the presence of Glu in position 3 of the recognition helices. The major accepted role of this amino acid residue is in discrimination rather than recognition, as discussed above (Figure 7), although in

Table 5. Hydrogen bonds involved in buttressing interactions between arginine and neighboring aspartate residues in Sp1 and EGR1 zinc fingers as monitored during the molecular dynamics simulations

	Sp1					EGR1			
	GAGGAGGAGC			**GGGGCGTGG**GG		**GCGTGGGCT**			
		Mean distance ± s.d	Mean angle ± s.d	Mean distance ± s.d	Mean angle ± s.d			Mean distance ± s.d	Mean angle ± s.d
Finger 3	Asp2 OD2 Arg-1 NH2	2.8 ± 0.1	157.0 ± 11.5	2.8 ± 0.1	160.2 ± 10.5	*Finger 3*	Asp2 OD2 Arg-1 NH2	2.8 ± 0.1	164.4 ± 8.0
RSDHLSK	Asp2 OD1 Arg-1 NE	2.8 ± 0.1	160.7 ± 9.9	2.8 ± 0.1	154.4 ± 4.4	**RSDERKR**	Asp2 OD1 Arg-1 NE	2.7 ± 0.1	158.2 ± 11.3
Finger 2	Asp2 OD2 Arg-1 NH2	2.8 ± 0.1	160.7 ± 9.9	2.8 ± 0.1	160.7 ± 9.9	*Finger 2*	Asp2 OD2 Arg-1 NH2	2.7 ± 0.1	165.5 ± 7.8
RSDELQR	Asp2 OD1 Arg-1 NE	2.8 ± 0.1	162.9 ± 8.9	2.9 ± 0.2	151.8 ± 14.4	**RSDHLTT**	Asp2 OD1 Arg-1 NE	2.8 ± 0.1	161.5 ± 9.7
						Finger 1	Asp2 OD2 Arg-1 NH2	2.8 ± 0.1	160.6 ± 10.5
						RSDELTR	Asp2 OD1 Arg-1 NE	2.8 ± 0.1	155.7 ± 12.4

Distances are in Å and angles are in degrees.

our simulation of the complex of EGR1 DBD with its binding site at the *mdr1* promoter, Glu3 in EGR1-f3 is consistently seen to hydrogen bond to the central cytosine base in the first GCG triplet (Table 2). Of note, this interaction is visible in the X-ray crystal structure of the complex of the Asp2(f1) → Ala point mutant of EGR1/Zif268 DBD with the sequence GCG TGG GCT (PDB code 1jk2).[77] Perhaps more importantly, in the complex of Sp1 DBD with its binding site at the *wt1* promoter, the long residence time and the proper orientation of a bound water molecule (Figure 8) suggest strongly that adenine in the middle position of the triplet can be recognized specifically by Glu3. Since this amino acid residue is present in EGR1-f1 and EGR1-f3, this finding can account for the good binding affinity of EGR1 to sites such as GAG TGG GAG and GAG GGG GAG.[20]

The role of His3 in sequence recognition appears to be the same in Sp1-f3 and in EGR1-f2 (which has been shown to bind to TAG equally well as to TGG),[20] both selecting for either G or A in the middle position of the subsite, whereas it is seen to depend on flanking sequences in the case of Sp1-f1, as discussed above. Overall, the similarity between EGR1-f2 and Sp1-f3 is slightly less than that between EGR1-f1/EGR1-f3 and Sp1-f2, which can be considered virtually equivalent.

Biological implications of GC-box occupancy by Sp1, EGR1 and WT1

GC-rich *cis*-regulatory sequences are present in the promoters of more than 1000 different genes involved in cell proliferation, differentiation and apoptosis. The transcription factors that bind to these sequences can behave as either activators or repressors through interactions with other members of the transcriptional machinery.[1] Since the prototypical representative is the ubiquitously expressed Sp1,[6] the promoter-binding sites are collectively termed Sp1 sites,[3] despite the fact that their different sequences must imply differences in their DNA-binding preferences.

Activation of the *mdr1* promoter by TPA and other agents increases transcription of the *mdr1* gene in a process that requires binding of EGR1 to the DNA sequence studied in this work.[32] The Wilms' tumor suppressor, WT1, another member of the EGR family, inhibits this TPA-induced response in K562 cells upon binding to this promoter, suggesting that the *mdr1* gene is a target for regulation by the *wt1* gene product.[78] Examination of the sequence of the four zinc fingers of this transcription factor (Figure 1) and consideration of their amino acid specificity according to the rules discussed above (Figure 6) suggest that WT1 can bind to the 5′-GCG TGG GCT GAG-3′ sequence overlapping the EGR1 binding sequence in the *mdr1* promoter (Figure 2) that is responsive to TPA, as well as the Sp1-binding site that is partly responsible for activation by UV-irradiation.[30]

Overlapping EGR1/Sp1/WT1 or EGR1/Sp1 sites similar to those described here for the *mdr1* promoter have been shown to be involved in the regulation of genes encoding, for example, adenosine deaminase (in which Sp1 acts as an activator and EGR1 acts as a repressor),[79] tyrosine hydroxylase,[80] human tumor necrosis factor,[81] tumor growth factor,[82] ornithine decarboxylase,[83] and platelet-derived growth factor A-chain (PDGF-A)[84] and B-chain (PDGF-B).[85] In this latter case, Sp1 has been shown to be bound to the promoter in unstimulated cells and to be displaced by EGR1 in a dose-dependent manner in response to vascular injury. Independent experiments have indeed shown that displacement of bound Sp1 by elevated amounts of EGR1 results in decreased Sp1-dependent transcription.[86] Therefore, in living cells, binding of one or another of these transcription factors to such overlapping sites will ultimately depend on their relative concentrations and their intrinsic affinities for their respective GC boxes. We have already proposed that the resulting balance in such competition could be influenced by the binding of small ligands such as ET743.[37,38]

Conclusions

To guide the construction of the zinc finger domains of Sp1 and their arrangement around their binding sites at the *mdr1* and *wt1* promoters, we have relied on previous studies on structure determination, site-directed mutagenesis, phage-display selections, and protection/interference data. Although a variety of docking arrangements has been observed in zinc finger-DNA complexes, the structure of the two Sp1–DNA complexes studied suggests a Zif268/EGR1-like docking arrangement, in accord with early suggestions,[61] with the capacity to recognize three consecutive DNA triplets through the concourse of the three recognition helices. Preferred sequences making up the first triplet (which is recognized by the third finger) are TGG, GGG and GAG. The third triplet (to which the first finger binds) can be TGG, GGG, GAG, GGA, or even GAA, whereas the second is more restrictive, in that only GCG or GAG are allowed.

Our simulations indicate that Sp1-f1 may not be so unusual as was initially thought. The main differences with respect to fingers 2 and 3 are the lack of Asp in position 2 (which makes direct recognition of the complementary C-strand outside the 9 nt box unlikely), the replacement of Arg with Lys in position −1 (which leads to larger side-chain fluctuations), and the presence of Ala in the position that should recognize the 5′ base in the third triplet subsite. This Ala in position +6 of Sp1-f1 can be considered as a non-DNA-contacting residue but its role may be important as a "permissive" amino acid that will allow the corresponding base site to be "read out" by the Asp2 in the following finger, which effectively

discriminates against A or C in the primary strand thereby specifying for either T or G, as is observed in all Sp1-binding sites.

The preceding considerations are supported by the available experimental evidence and the parallel results obtained from the simulation of a "control" EGR1–DNA complex. The data presented strongly indicate the need to extend the range of sequences that can be recognized by Sp1 (and possibly other members of the Sp1-like family of transcription factors as well) beyond the commonly accepted "minimal" GGGCGG hexanucleotide core. This sequence (which is composed of differently sized fragments of three triplet subsites) does not include, for example, the TGG GCG TGG sequence recently selected as a high-affinity site,[68] the (inverted) GAG GCG GAG present in the *hTERT* promoter,[67] or the GAG (= CTC) repeat that has been shown conclusively to be a high-affinity site for Sp1 binding to the *wt1* promoter,[18] in agreement with the stability of the complex studied herein. For this reason, the names GC-box or GC-element appear to be inadequate or misleading for describing the canonical-binding site of Sp1, as they make reference to only a small subset of sites or to just one particular core consensus sequence. Consequently, searches for putative Sp1-binding sites defined on the basis of an inappropriate consensus sequence made up of a rather arbitrary collection of G and C bases are likely to yield incorrect or inaccurate results. We would favor a definition of binding sites in terms of a juxtaposition of subsites that take into account the selectivity rules derived for zinc fingers (both in terms of accepted and excluded bases), as shown in Figure 6 for Sp1 and EGR1. It is then clear that alternative binding sequences need to be explored further, as they may have passed unnoticed. In addition, it must be borne in mind that the basal affinity provided by the zinc fingers can be surpassed by other protein domains, as shown for EGR1 family members[20] and, as suggested by one of the reviewers, will most probably be co-dependent on other transcription factors that act in concert with Sp1.

The approach described here to model the zinc finger domains of Sp1 can be extended easily to other Sp1-like transcription factors. The ensuing structure-based results can aid in the assignment of particular DNA stretches as putative-binding sites leading to fairly accurate predictions and a reduction in the number of experiments that need to be performed.

Methodology

Construction and refinement of the starting DNA–protein complexes

The Zif268(DBD)-DNA crystal structure at 1.6 Å resolution (PDB entry 1aay)[35] was used to model the EGR1(DBD)–DNA complex after introduction of the appropriate modifications of base composition in the DNA template. The complex was then refined by means of 2000 steps of steepest descent energy minimization keeping all the atoms fixed to their initial positions except those belonging to the replaced residues, which were free to move. The second-generation AMBER force-field,[87] including parameters for Zn^{2+} and incorporating new parameters for improved sugar pucker phases and helical repeat (parm99),[88] was employed throughout.

To build the initial model for the Sp1–DNA complexes, the automated comparative protein modeling server SWISS-MODEL[89] was used and the five X-ray crystal structures of DNA-bound native and mutant EGR1/Zif268 DBD (PDB codes 1meyC, 1meyF, 1a1f, 1a1g, and 1jk1) were chosen as templates. The resulting protein model was mounted onto one of the EGR1–DNA complexes (PDB 1a1f) using the corresponding C^{α} atoms in the superimposition. This was followed by appropriate changes in base composition to model the two DNA sequences, extension of the DNA helix using standard *B*-DNA parameters[90] to provide the 14-mers, and by a manual search for optimal side-chain conformations. For each mutated residue, the rotamer that did not give rise to any steric clash with neighboring atoms and at the same time led to better agreement with the experimental data was initially chosen. A short optimization run restraining all non-H atoms in the non-mutated residues to their initial coordinates then allowed readjustment of covalent bonds and van der Waals contacts without changing the overall conformation of the complexes.

Molecular dynamics of DNA-bound EGR1 and Sp1 in aqueous solution

Each molecular system was neutralized by addition of the appropriate number of sodium ions[91] (12 and nine for the Sp1 and EGR1 systems, respectively), placed in electrostatically favored positions, and immersed in a rectangular box of ~7500 TIP3P water molecules.[92] Each water box extended 7 Å away from any solute atom, and the cutoff distance for the non-bonded interactions was 9 Å. Periodic boundary conditions were applied and electrostatic interactions were represented using the smooth particle mesh Ewald method[93] with a grid spacing of ~ 1 Å. Unrestrained molecular dynamics simulations at 300 K and 1 atm (= 101,325 Pa) were then run for 3 ns using the SANDER module in AMBER6.0†. The coupling constants for the temperature and the pressure baths were 1.0 ps and 0.2 ps, respectively. SHAKE[94] was applied to all bonds involving hydrogen atoms and an integration step of 2 fs was used throughout. The simulation protocol was essentially as described,[36,38] involving a series of progressive energy minimizations followed by a 20 ps heating phase and a 70 ps equilibration period before data collection. System coordinates were saved every 2 ps after the first 500 ps of equilibration for further analysis.

Analysis of the molecular dynamics trajectories

Three-dimensional structures and trajectories were visually inspected using the computer graphics program Insight II (Insight II, version 98.0, Molecular Simulations Inc. 9685 Scranton Road, San Diego, CA 92121-2777,

† http://www.amber.ucsf.edu/amber/amber.html

USA). Root-mean-square (rms) deviations from the initial structures, interatomic distances and distribution functions were monitored using the CARNAL module in AMBER. The conformational and helical parameters of the DNA molecules were analyzed by means of programmes CURVES[95] and FREEHELIX.[96]

All calculations were performed on the SGI R8000 Power Challenge at Alcalá University Computer Center, on the SGI R14000 Origin 3800 at CIEMAT (Madrid), and locally on SGI R5000 O2 workstations.

Acknowledgements

E.M. is the recipient of a research fellowship from Junta de Comunidades de Castilla-La Mancha. A studentship to R. G.-N. and financial support from Pharma-Mar S.A. (Tres Cantos, Madrid) are gratefully acknowledged. We thank the University of Alcalá Computing Centre and the CIEMAT (Madrid) for generous allowances of computer time on their SGI servers. This research was supported, in part, by the National Foundation for Cancer Research.

References

1. Philipsen, S. & Suske, G. (1999). A tale of three fingers: the family of mammalian Sp/XKLF transcription factors. *Nucl. Acids Res.* **27**, 2991–3000.
2. Suske, G. (1999). The Sp-family of transcription factors. *Gene*, **238**, 291–300.
3. Black, A. R., Black, J. D. & Azizkhan-Clifford, J. (2001). Sp1 and krüppel-like factor family of transcription factors in cell growth regulation and cancer. *J. Cell. Physiol.* **188**, 143–160.
4. Choo, Y. & Klug, A. (1997). Physical basis of a protein-DNA recognition code. *Curr. Opin. Struct. Biol.* **7**, 117–125.
5. Wolfe, S. A., Greisman, H. A., Ramm, E. I. & Pabo, C. O. (1999). Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *J. Mol. Biol.* **285**, 1917–1934.
6. Kadonaga, J. T., Jones, K. A. & Tjian, R. (1986). Promoter specific activation of RNA polymerase II by Sp1. *Trends Biochem. Sci.* **11**, 20–23.
7. Liu, C., Calogero, A., Ragona, G., Adamson, E. & Mercola, D. (1996). EGR-1, the reluctant suppression factor. *Crit. Rev. Oncol.* **7**, 101–125.
8. Christy, B. & Nathans, D. (1989). DNA binding site of the growth factor-inducible protein Zif268. *Proc. Natl Acad. Sci. USA*, **86**, 8737–8741.
9. Khachigian, L. M., Lindner, V., Williams, A. J. & Collins, T. (1996). Egr-1-induced endothelial gene expression: a common theme in vascular injury. *Science*, **271**, 1427–1431.
10. Matera, A. G. & Ward, D. C. (1993). Localization of the human Sp1 transcription factor gene to 12q13 by fluorescence *in situ* hybridization. *Genomics*, **17**, 793–794.
11. Le Beau, M. M., Espinosa, R., III, Neuman, W. L., Stock, W., Roulston, D., Larson, R. A. *et al.* (1993). Cytogenetic and molecular delineation of the smallest commonly deleted region of chromosome 5 in malignant myeloid diseases. *Proc. Natl Acad. Sci. USA*, **90**, 5484–5488.
12. Horrigan, S. K., Arbieva, Z. H., Xie, H. Y., Kravarusic, J., Fulton, N. C., Naik, H. *et al.* (2000). Delineation of a minimal interval and identification of 9 candidates for a tumor suppressor gene in malignant myeloid disorders on 5q31. *Blood*, **95**, 2372–2377.
13. Huang, R-P., Liu, C., Fan, Y., Mercola, D. & Adamson, E. (1995). Egr-1 negatively regulates human tumor cell growth *via* the DNA-binding domain. *Cancer Res.* **55**, 5054–5062.
14. Liu, C., Adamson, E. & Mercola, D. (1996). Transcription factor EGR-1 suppresses the growth and transformation of human HT-1080 fibrosarcoma cells by induction of transforming growth factor β1. *Proc. Natl Acad. Sci. USA*, **93**, 11831–11836.
15. Rauscher, F. J., III (1993). The *WT1* Wilms tumor gene product: a developmentally regulated transcription factor in the kidney that functions as a tumor suppressor. *FASEB J.* **7**, 896–903.
16. Nakagama, H., Heinrich, G., Pelletier, J. & Housman, D. E. (1995). Sequence and structural requirements for high-affinity DNA binding by the WT1 gene product. *Mol. Cell. Biol.* **15**, 1489–1498.
17. Scharnhorst, V., Dekker, P., van der Eb, A. J. & Jochemsen, A. G. (2000). Physical interaction between Wilms tumor 1 and p73 proteins modulates their functions. *J. Biol. Chem.* **275**, 10202–10211.
18. Cohen, H. T., Bossone, S. A., Zhu, G., McDonald, G. A. & Sukhatme, V. P. (1997). Sp1 is a critical regulator of the Wilms' tumor-1 gene. *J. Biol. Chem.* **272**, 2901–2913.
19. Wang, Z., Qiu, Q., Enger, K. T. & Deuel, T. F. (1993). A second transcriptionally active DNA-binding site for the Wilms tumor gene product, WT1. *Proc. Natl Acad. Sci. USA*, **90**, 8896–8900.
20. Swirnoff, A. H. & Milbrandt, J. (1995). DNA-binding specificity of NGFI-A and related zinc finger transcription factors. *Mol. Cell. Biol.* **15**, 2275–2287.
21. Madden, M. J., Morrow, C. S., Nakagawa, M., Goldsmith, M. E., Fairchild, C. R. & Cowan, K. H. (1993). Identification of 5′ and 3′ sequences involved in the regulation of transcription of the human *mdr1* gene *in vivo*. *J. Biol. Chem.* **268**, 8290–8297.
22. Ambudkar, S. V., Dey, S., Hrycyna, C. A., Ramachandra, M., Pastan, I. & Gottesman, M. M. (1999). Biochemical, cellular, and pharmacological aspects of the multidrug transporter. *Annu. Rev. Pharmacol. Toxicol.* **39**, 361–398.
23. Abolhoda, A., Wilson, A. E., Ross, H., Danenberg, P. V., Burt, M. & Scotto, K. W. (1999). Rapid activation of MDR1 gene expression in human metastatic sarcoma after *in vivo* exposure to doxorubicin. *Clin. Cancer Res.* **5**, 3352–3356.
24. Lehne, G. (2000). P-glycoprotein as a drug target in the treatment of multidrug resistant cancer. *Curr. Drug Targets*, **1**, 85–99.
25. Jin, S., Gorfajn, B., Faircloth, G. & Scotto, K. W. (2000). Ecteinascidin 743, a transcription-targeted chemotherapeutic that inhibits *MDR1* activation. *Proc. Natl Acad. Sci. USA*, **97**, 6775–6779.
26. Scotto, K. W. & Johnson, R. A. (2001). Transcription of the multidrug resistance gene MDR1: a therapeutic target. *Mol. Interventions*, **1**, 14–22.
27. Bartsevich, V. V. & Juliano, R. L. (2000). Regulation of the MDR1 gene by transcriptional repressors selected using peptide combinatorial libraries. *Mol. Pharmacol.* **58**, 1–10.

28. Manzanares, I., Cuevas, C., García-Nieto, R., Marco, E. & Gago, F. (2001). Advances in the chemistry and pharmacology of ecteinascidins, a promising new class of anticancer agents. *Curr. Med. Chem.-Anti-Cancer Agents*, **1**, 257–276.
29. Sundseth, R., MacDonald, G., Ting, J. & King, A. C. (1997). DNA elements recognizing NF-Y and Sp1 regulate the human multidrug-resistance gene promoter. *Mol. Pharmacol.* **51**, 963–971.
30. Hu, Z., Jin, S. & Scotto, K. W. (2000). Transcriptional activation of the MDR1 gene by UV irradiation. Role of NF-Y and Sp1. *J. Biol. Chem.* **275**, 2979–2985.
31. Cornwell, M. M. & Smith, D. E. (1993). Sp1 activates the *MDR1* promoter through one of two distinct G-rich regions that modulate promoter activity. *J. Biol. Chem.* **268**, 19505–19511.
32. McCoy, C., Smith, D. E. & Cornwell, M. M. (1995). 12-*O*-tetradecanoylphorbol-13-acetate activation of the MDR1 promoter is mediated by EGR1. *Mol. Cell. Biol.* **15**, 6100–6108.
33. McCoy, C., McGee, S. B. & Cornwell, M. M. (1999). The Wilms' tumor suppressor, WT1, inhibits 12-*O*-tetradecanoylphorbol-13-acetate activation of the multidrug resistance-1 promoter. *Cell Growth Differ.* **10**, 377–386.
34. Nekludova, L. & Pabo, C. O. (1994). Distinctive DNA conformation with enlarged major groove is found in Zn-finger–DNA and other protein–DNA complexes. *Proc. Natl Acad. Sci. USA*, **91**, 6948–6952.
35. Pavletich, N. P. & Pabo, C. O. (1991). Zinc finger–DNA recognition: crystal structure of a Zif268–DNA complex at 2.1 Å. *Science*, **252**, 809–817.
36. García-Nieto, R., Manzanares, I., Cuevas, C. & Gago, F. (2000). Bending of DNA upon binding of ecteinascidin 743 and phthalascidin 650 studied by unrestrained molecular dynamics simulations. *J. Am. Chem. Soc.* **122**, 7172–7182.
37. García-Nieto, R., Manzanares, I., Cuevas, C. & Gago, F. (2000). Increased DNA binding specificity for anti-tumor ecteinascidin ET743 through protein–DNA interactions? *J. Med. Chem.* **43**, 4367–4369.
38. Marco, E., García-Nieto, R., Mendieta, J., Manzanares, I., Cuevas, C. & Gago, F. (2002). A 3·(ET743)–DNA complex that both resembles an RNA–DNA hybrid and mimics zinc finger-induced DNA structural distortions. *J. Med. Chem.* **45**, 871–880.
39. Norberg, J. & Nilsson, L. (2002). Molecular dynamics applied to nucleic acids. *Acc. Chem. Res.* **35**, 465–472.
40. Cheatham, T. E., III & Kollman, P. A. (2000). Molecular dynamics simulation of nucleic acids. *Annu. Rev. Phys. Chem.* **51**, 435–471.
41. Narayan, V. A., Kriwacki, R. W. & Caradonna, J. P. (1997). Structures of zinc finger domains from transcription factor Sp1. Insights into sequence-specific protein–DNA recognition. *J. Biol. Chem.* **272**, 7801–7809.
42. Wolfe, S. A., Nekludova, L. & Pabo, C. O. (2000). DNA recognition by Cys_2His_2 zinc finger proteins. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 183–212.
43. Desjarlais, J. R. & Berg, J. M. (1992). Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proc. Natl Acad. Sci. USA*, **89**, 7345–7349.
44. Jamieson, A. C., Wang, H. & Kim, S. H. (1996). A zinc finger directory for high-affinity DNA recognition. *Proc. Natl Acad. Sci. USA*, **93**, 12834–12839.
45. Segal, D. J., Dreier, B., Beerli, R. R. & Barbas, C. F., III (1999). Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5′-GNN-3′ DNA target sequences. *Proc. Natl Acad. Sci. USA*, **96**, 2758–2763.
46. Isalan, M., Choo, Y. & Klug, A. (1997). Synergy between adjacent zinc fingers in sequence-specific DNA recognition. *Proc. Natl Acad. Sci. USA*, **94**, 5617–5621.
47. Dreier, B., Segal, D. J. & Barbas, C. F., III (2000). Insights into the molecular recognition of the 5′-GNN-3′ family of DNA sequences by zinc finger domains. *J. Mol. Biol.* **303**, 489–502.
48. Segal, D. J. & Barbas, C. F., III (2000). Design of novel sequence-specific DNA-binding proteins. *Curr. Opin. Chem. Biol.* **4**, 34–39.
49. Choo, Y. & Isalan, M. (2000). Advances in zinc finger engineering. *Curr. Opin. Struct. Biol.* **10**, 411–416.
50. Beerli, R. R., Segal, D. J., Dreier, B. & Barbas, C. F., III (1998). Toward controlling gene expression at will: specific regulation of the erbB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks. *Proc. Natl Acad. Sci. USA*, **95**, 14628–14633.
51. Kuwahara, J., Yonezawa, A., Futamura, M. & Sugiura, Y. (1993). Binding of transcription factor Sp1 to GC box DNA revealed by footprinting analysis: different contact of three zinc fingers and sequence recognition mode. *Biochemistry*, **32**, 5994–6001.
52. Gidoni, D., Dynan, W. S. & Tjian, R. (1984). Multiple specific contacts between a mammalian transcription factor and its cognate promoters. *Nature*, **312**, 409–413.
53. Nagaoka, M. & Sugiura, Y. (1996). Distinct phosphate backbone contacts revealed by some mutant peptides of zinc finger protein Sp1: effect of protein-induced bending on DNA recognition. *Biochemistry*, **35**, 8761–8768.
54. Letovsky, J. & Dynan, W. S. (1989). Measurement of the binding of transcription factor Sp1 to a single GC box recognition sequence. *Nucl. Acids Res.* **17**, 2639–2653.
55. Thiesen, H. J. & Bach, C. (1990). Target detection assay (TDA): a versatile procedure to determine DNA binding sites as demonstrated on SP1 protein. *Nucl. Acids Res.* **18**, 3203–3209.
56. Kriwacki, R. W., Schultz, S. C., Steitz, T. A. & Caradonna, J. P. (1992). Sequence-specific recognition of DNA by zinc-finger peptides derived from the transcription factor Sp1. *Proc. Natl Acad. Sci. USA*, **89**, 9759–9763.
57. Thiesen, H. J. & Schroder, B. (1991). Amino acid substitutions in the SP1 zinc finger domain alter the DNA binding affinity to cognate SP1 target site. *Biochem. Biophys. Res. Commun.* **175**, 333–338.
58. Sjøttem, E., Andersen, C. & Johansen, T. (1997). Structural and functional analyses of DNA bending induced by Sp1 family transcription factors. *J. Mol. Biol.* **267**, 490–504.
59. Yokono, M., Saegusa, N., Matsushita, K. & Sugiura, Y. (1998). Unique DNA binding mode of the N-terminal zinc finger of transcription factor Sp1. *Biochemistry*, **37**, 6824–6832.
60. Rotheneder, H., Geymayer, S. & Haidweger, E. (1999). Transcription factors of the Sp1 family: interaction with E2F and regulation of the murine thymidine kinase promoter. *J. Mol. Biol.* **293**, 1005–1015.
61. Berg, J. M. (1992). Sp1 and the subfamily of zinc finger proteins with guanine-rich binding sites. *Proc. Natl Acad. Sci. USA*, **89**, 11109–11110.

62. Nardelli, J., Gibson, T. J., Vesque, C. & Charnay, P. (1991). Base sequence discrimination by zinc-finger DNA-binding domains. *Nature*, **349**, 175–178.
63. Desjarlais, J. R. & Berg, J. M. (1993). Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins. *Proc. Natl Acad. Sci. USA*, **90**, 2256–2260.
64. Nagaoka, M., Shiraishi, Y., Uno, Y., Nomura, W. & Sugiura, Y. (2002). Interconversion between serine and aspartic acid in the alpha helix of the n-terminal zinc finger of sp1: implication for general recognition code and for design of novel zinc finger peptide recognizing complementary strand. *Biochemistry*, **41**, 8819–8825.
65. Choo, Y. (1998). End effects in DNA recognition by zinc finger arrays. *Nucl. Acids Res.* **26**, 554–557.
66. Thukral, S. K., Morrison, M. L. & Young, E. T. (1991). Alanine scanning site-directed mutagenesis of the zinc fingers of transcription factor ADR1: residues that contact DNA and that transactivate. *Proc. Natl Acad. Sci. USA*, **88**, 9188–9192.
67. Won, J., Yim, J. & Kim, T. K. (2002). Sp1 and Sp3 recruit histone deacetylase to repress transcription of human telomerase reverse transcriptase (hTERT) promoter in normal human somatic cells. *J. Biol. Chem.* **277**, 38230–38238.
68. Nagaoka, M., Shiraishi, Y. & Sugiura, Y. (2001). Selected base sequence outside the target binding site of zinc finger protein Sp1. *Nucl. Acid Res.* **29**, 4920–4929.
69. Ciudad, C. J., Morris, A. E., Jeng, C. & Chasin, L. A. (1992). Point mutational analysis of the hamster dihydrofolate reductase minimum promoter. *J. Biol. Chem.* **267**, 3650–3656.
70. Packer, M. J., Dauncey, M. P. & Hunter, C. A. (2000). Sequence-dependent DNA structure: tetranucleotide conformational maps. *J. Mol. Biol.* **295**, 85–103.
71. Uno, Y., Matsushita, K., Nagaoka, M. & Sugiura, Y. (2001). Finger-positional change in three zinc finger protein Sp1: influence of terminal finger in DNA recognition. *Biochemistry*, **40**, 1787–1795.
72. Bowers, P. M., Schaufler, L. E. & Klevit, R. E. (1999). A folding transition and novel zinc finger accessory domain in the transcription factor ADR1. *Nature Struct. Biol.* **6**, 478–485.
73. Sun, D. & Hurley, L. H. (1994). Cooperative bending of the 21-base-pair repeats of the SV40 viral early promoter by human Sp1. *Biochemistry*, **33**, 9578–9587.
74. Shi, Y. & Berg, J. M. (1996). DNA unwinding induced by zinc finger protein binding. *Biochemistry*, **35**, 3845–3848.
75. Ikeda, K., Nagano, K. & Kawakami, K. (1993). Possible implications of Sp1-induced bending of DNA on synergistic activation of transcription. *Gene*, **136**, 341–343.
76. Mastrangelo, I. A., Courey, A. J., Wall, J. S., Jackson, S. P. & Hough, P. V. (1991). DNA looping and Sp1 multimer links: a mechanism for transcriptional synergism and enhancement. *Proc. Natl Acad. Sci. USA*, **88**, 5670–5674.
77. Miller, J. C. & Pabo, C. O. (2001). Rearrangement of side-chains in a Zif268 mutant highlights the complexities of zinc finger-DNA recognition. *J. Mol. Biol.* **313**, 309–315.
78. McCoy, C., McGee, S. B. & Cornwell, M. M. (1999). The Wilms' tumor suppressor, WT1, inhibits 12-*O*-tetradecanoylphorbol-13-acetate activation of the multidrug resistance-1 promoter. *Cell Growth Differ.* **10**, 377–386.
79. Ackerman, S. L., Minden, A. G., Williams, G. T., Bobonis, C. & Yeung, C. Y. (1991). Functional significance of an overlapping consensus binding motif for Sp1 and Zif268 in the murine adenosine deaminase gene promoter. *Proc. Natl Acad. Sci. USA*, **88**, 7523–7527.
80. Papanikolaou, N. A. & Sabban, E. L. (2000). Ability of Egr1 to activate tyrosine hydroxylase transcription in PC12 cells. Cross-talk with AP-1 factors. *J. Biol. Chem.* **275**, 26683–26689.
81. Kramer, B., Meichle, A., Hensel, G., Charnay, P. & Kronke, M. (1994). Characterization of an Krox-24/Egr-1-responsive element in the human tumor necrosis factor promoter. *Biochim. Biophys. Acta*, **1219**, 413–421.
82. Udvadia, A. J., Rogers, K. T. & Horowitz, J. M. (1992). A common set of nuclear factors bind to promoter elements regulated by the retinoblastoma protein. *Cell Growth Differ.* **3**, 597–608.
83. Moshier, J. A., Skunca, M., Wu, W., Boppana, S. M., Rauscher, F. J., III & Dosescu, J. (1996). Regulation of ornithine decarboxylase gene expression by the Wilms' tumor suppressor WT1. *Nucl. Acids Res.* **24**, 1149–1157.
84. Khachigian, L. M., Williams, A. J. & Collins, T. (1995). Interplay of Sp1 and Egr-1 in the proximal platelet-derived growth factor A-chain promoter in cultured vascular endothelial cells. *J. Biol. Chem.* **270**, 27679–27686.
85. Khachigian, L. M., Lindner, V., Williams, A. J. & Collins, T. (1996). Egr-1-induced endothelial gene expression: a common theme in vascular injury. *Science*, **271**, 1427–1431.
86. Liu, C., Calogero, A., Ragona, G., Adamson, E. & Mercola, D. (1996). EGR-1, the reluctant suppression factor. *Crit. Rev. Oncol.* **7**, 101–125.
87. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M. *et al.* (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197.
88. Cheatham, T. E., Cieplak, P. & Kollman, P. A. (1999). A modified version of the Cornell *et al*. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dynam.* **16**, 845–862.
89. Guex, N., Diemand, A. & Peitsch, M. C. (1999). Protein modelling for all. *Trends Biochem. Sci.* **24**, 364–367.
90. Arnott, S. & Hukins, D. W. (1972). Optimised parameters for A-DNA and B-DNA. *Biochem. Biophys. Res. Commun.* **47**, 1504–1509.
91. Åqvist, J. (1990). Ion–water interaction potentials derived from free energy perturbation simulations. *J. Phys. Chem.* **94**, 8021–8024.
92. Jorgensen, W. L., Chandrasekhar, J. & Madura, J. D. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935.
93. Darden, T. A., York, D. & Pedersen, L. G. (1993). Particle mesh Ewald: an N*log(N) method for computing Ewald sums. *J. Chem. Phys.* **98**, 10089–10092.
94. Ryckaert, J. P., Ciccoti, G. & Berendsen, H. J. C. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341.
95. Lavery, R. & Sklenar, H. (1988). The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dynam.* **6**, 63–91.

96. Dickerson, R. E. (1998). DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucl. Acids Res.* **26**, 1906–1926.
97. Zhong, Z.-D., Hammani, K., Bae, W. S. & DeClerck, Y. A. (2000). NF-Y and Sp1 cooperate for the transcriptional activation and cAMP response of human tissue inhibitor of metalloproteinases-2. *J. Biol. Chem.* **275**, 18602–18610.
98. Nicolás, M., Noé, V., Jensen, K. B. & Ciudad, C. J. (2001). Cloning and characterization of the 5′-flanking region of the human transcription factor *Sp1* gene. *J. Biol. Chem.* **276**, 22126–22132.

Edited by B. Honig

(Received 9 September 2002; received in revised form 14 February 2003; accepted 14 February 2003)