Chapter 2

INTRODUCTION

Protein/DNA Interactions

The building plan of living organisms is stored as a unique sequence of the four bases adenine, guanine, cytosine and thymine in a macromolecule named deoxyribonucleic acid (DNA). This heritable information is selectively read out through specific protein/DNA interactions in a process called transcription. After transcription, the information is translated into a multitude of functionally distinct proteins. The sophisticated assembly and fine-tuned function of these proteins give rise to unique the phenotypes of living individuals. Thus, DNA binding proteins involved in the initiation, regulation and termination of DNA read-out stand at the start of the processes which determine the morphology and function of cells, tissues and whole organisms.

During the last two decades, marked advancements in the biosynthetic production and high resolution structural analysis of proteins and DNA have allowed the revelation of a broad structural basis for protein/DNA interactions. Transcription factors belong to the best studied group of proteins involved in DNA read-out. They commonly bind to certain consensus sequences in the promoter region of a gene, thereby activating the transcription of that gene through additional interactions with the transcription complex. On the contrary, antagonists of activating transcription factors, called transcriptional repressors, also bind to certain consensus sequences in regulatory regions of a gene, thereby repressing its transcription. Both activating and repressing transcription factors commonly contain a DNA binding domain of 60 – 90 residues, which belong to either of the five classes of protein folds: helix-turn-helix, Zn-binding, leucine zipper, β-sheet and immunoglobulin-like folds.

Helix-turn-helix DNA binding proteins

Helix-turn-helix (HTH) DNA binding protein domains belong to the most thoroughly studied class of DNA binding proteins. A large number of prokaryotic transcriptional regulators and eukaryotic homeodomains, many of which are also involved in transcriptional regulation, fall into this class. Most prokaryotic HTH domains bind cognate DNA as homodimers, or in some cases as homotetramers. In contrast, eukaryotic HTH domains often bind DNA as monomers utilizing N- or C-terminal arms to mediate additional DNA contacts [29]. The two α -helices of the HTH motif, which cross at an angle of approximately 120°, are connected by a turn of about three residues. HTH variants in which this turn is extended to a longer loop are frequently named helix-loop-helix domains instead of HTH domains. Since two α-helices alone are insufficient to enclose a hydrophobic core necessary for folding, HTH domains generally contain a third α-helix N-terminal to the HTH motif.

When bound to DNA, the C-terminal α-helix is inserted into the major groove of B-DNA where it mediates the majority of specific DNA interactions. Therefore, this α-helix is referred to as ‘recognition helix’. The solution structure of the *Antennapedia* homeodomain complexed with cognate DNA [30] shows that α-helix 3, the recognition helix, penetrates deeply into the major groove with its α-helical axis roughly parallel the groove (fig. 3). The flexible N-terminal arm mediates additional DNA contacts in the minor groove. Both the recognition helix and the N-terminal arm form base-specific interactions. The loop between α-helix 1 and 2, and the N-terminal residues of α-helix 2 contact the DNA backbone. The DNA maintains an undistorted B-DNA conformation in complex with the *Antennapedia* homeodomain. Similar DNA binding geometries were observed for the homeodomains POU [31], *engrailed* [32], *even-skipped* [33] and MATα2 [34]. In comparison with eukaryotic homeodomains, the orientation of the recognition helix with respect to the DNA is more variant in prokaryotic HTH domains. Taken together, all HTH domains insert an α-helix into the major groove of B-DNA for sequence specific DNA recognition. The major groove is perfectly

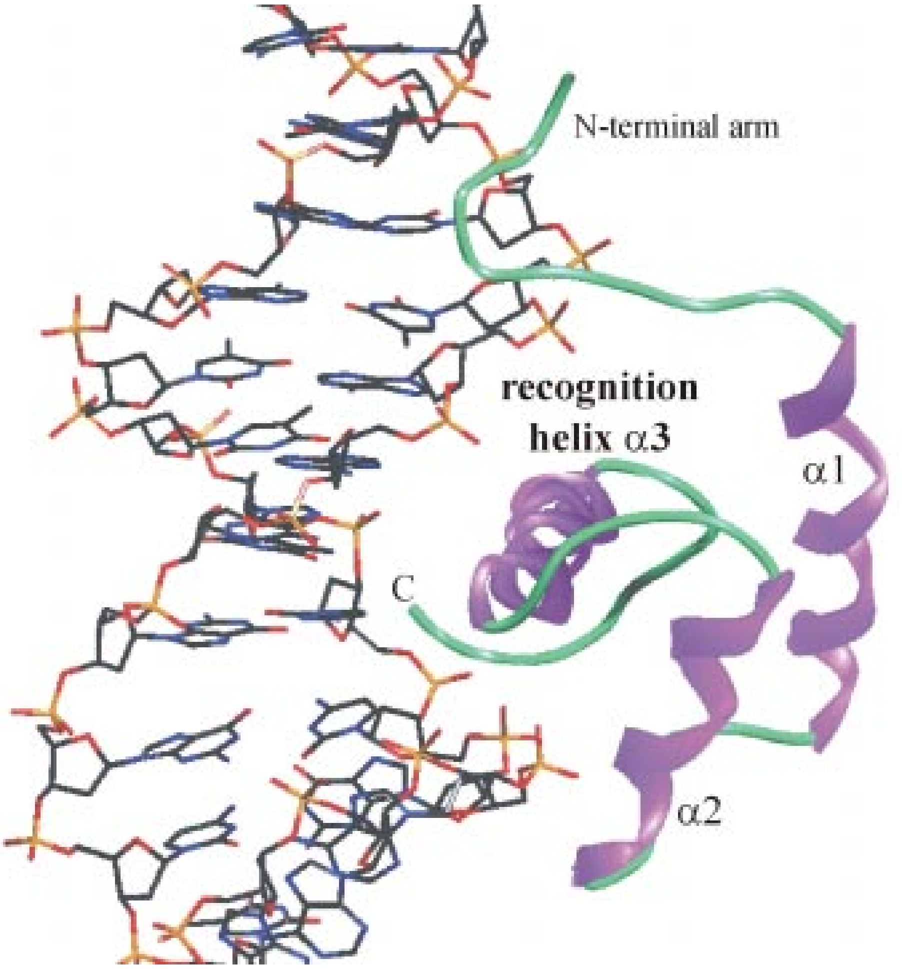

fig. 3 **The NMR structure of the *Antennapedia* homeodomain complexed with cognate DNA** [30] shows how HTH domains penetrate the major groove with their recognition helix α3 to mediate base specific DNA contacts. An N-terminal arm, which is rich in positively charged residues, forms additional DNA contacts in the minor groove.

shaped to accommodate one half turn of an α-helix which delivers a firm scaffold that presents more flexible amino acid side chains to the bases [35]. The other half turn of such a recognition helix is generally rich in hydrophobic residues that pack the α-helix to the protein domain. This simple and thermodynamically stable arrangement has been highly successful during evolution, as evidenced by the large number of prokaryotic and eukaryotic DNA binding domains using a recognition helix.

(α+β)HTH DNA binding proteins

In (α+β)HTH domains, the basic three-helix HTH fold is diversified by in the insertion of additional β-strands N- and C-terminal to the HTH motif. The crystal structure of heptatocyte nuclear factor 3γ (HNF-3γ) [36] shows that the C-terminal β-strands form an antiparallel β-sheet that is folded against the α-helical core (fig. 4). This increases the enclosed hydrophobic core of the domain. The β-strand preceding helix 2 contacts β3 of the C-terminal

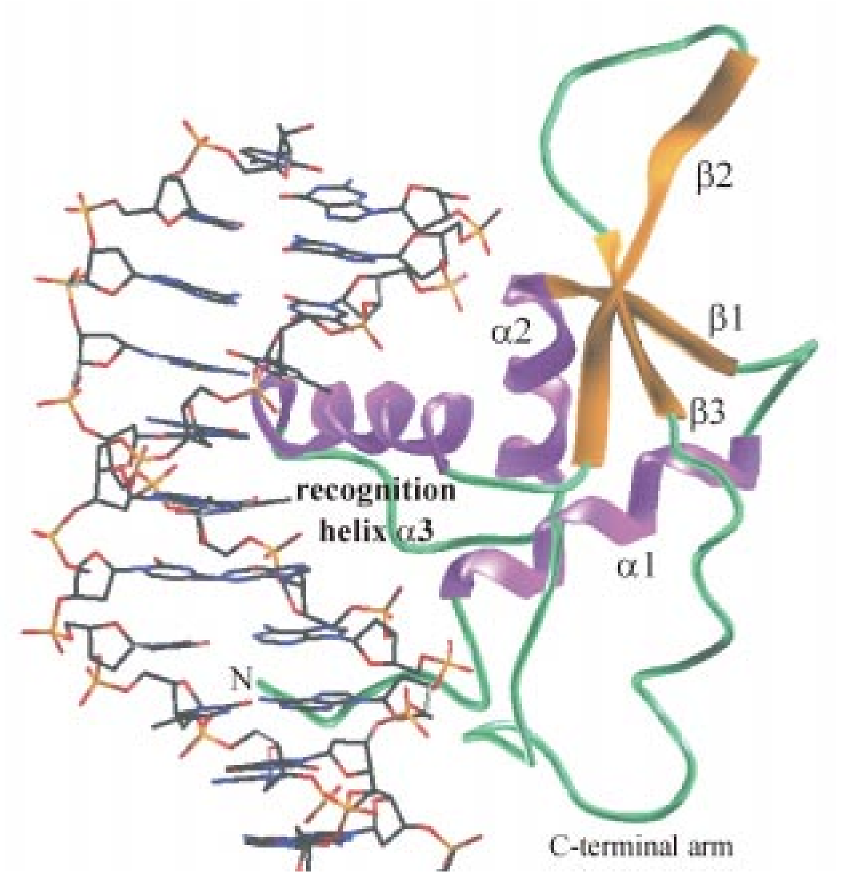

fig. 4 (α+β)HTH domains possess additional β-strands embedded N- and C-terminal to the HTH-motif but retain the pivotal recognition-helix/major-groove interaction. The crystal structure of the heptatocyte nuclear factor 3γ (HNF-3γ) complexed with cognate DNA [36], a eukaryotic member of the (α+β)HTH family, shows that one β-strand precedes helix 2 and that an antiparallel β-sheet succeeds recognition helix α3. The recognition helix and the N-termini of helix 1 and 2 form DNA contacts which are supplemented by contacts through the C-terminal β-sheet and a flexible C-terminal arm which contains a cluster of lysine and arginine residues.

β-sheet. Furthermore, the C-terminal β-sheet forms additional DNA contacts. In conjunction with the recognition helix, the C-terminal β-sheet locks the (α+β)HTH domain in a certain DNA binding geometry. A flexible arm at the C-terminus, which is rich in positively charged lysine and arginine residues, mediates additional DNA contacts. These C-terminal DNA contacts may be important for high-affinity binding by HNF-3γ compensating for the smaller interaction surface of the monomeric HNF-3γ/DNA complex, as compared with dimeric (HTH-domain)$_2$/DNA complexes. Further DNA contacting residues reside at the N-termini of helix 1 and 2. A similar arrangement of DNA contacting residues is found in other (α+β)HTH domains, such as catabolite activator protein (CAP) [37], the cell cycle transcription factor E2F-4 [38] and diphtheria toxin repressor (DtxR) [39]. In summary, (α+β)HTH DNA binding proteins demonstrate that the basic HTH motif can be diversified through β-strands at both ends to obtain modified DNA binding properties while retaining the fundamental recognition-helix/major groove contact.

Zinc binding domains that bind to DNA

This class of DNA binding domains, which contains a Zn^{2+} as a structural element, can be subdivided into three structurally distinct groups [40]. The first group comprises the canonical Zn-finger domains, which span approximately 30 residues and ligate one Zn^{2+} by two cysteines and two histidines. The second group encompasses approximately 70-residues domains found in steroid and related hormone receptors. These domains coordinate two Zn^{2+} cations by four cysteines each. The third group found in a set of yeast activators, including GAL4, packs two Zn^{2+} cations side by side liganded by six cysteines.

The first group, the Zn-fingers, is the most abundant with more than 1000 domains having been identified in various eukaryotes. Their sequence conforms to the following consensus:

$$hxCx_{2\text{-}4}Cx_3hx_5hx_2Hx_{3\text{-}4}Hx_{2\text{-}6}$$

where h is a hydrophobic residue, x is variable and the two cysteines (C) and histidines (H) are invariant. Their three-dimensional structure is characterized by an antiparallel β-sheet followed by a 12-residue α-helix. The two cysteines flank the β-turn, and the two histidines reside on the inward-facing side of the α-helix. Thus, coordination of the common Zn^{2+} packs the α-helix tightly against the β-sheet forming a remarkably stable ββα module. Most Zn-finger proteins contain three or more ββα modules in a line linked by a few flexible residues.

The crystal structure of the Zn-finger protein Zif268 complexed with cognate DNA demonstrates that three linked ββα modules bind to DNA by inserting their consecutive α-helices into the major groove in basically identical manner (fig. 5) [41]. This mode of DNA recognition resembles fingers that are inserted into adjacent major grooves giving rise to the name Zn-finger. Each finger contacts a block of three basepairs without a gap between consecutive blocks. Thus, the three modules of Zif268 read out nine contiguous basepairs, thereby allowing for high affinity, sequence specific DNA recognition.

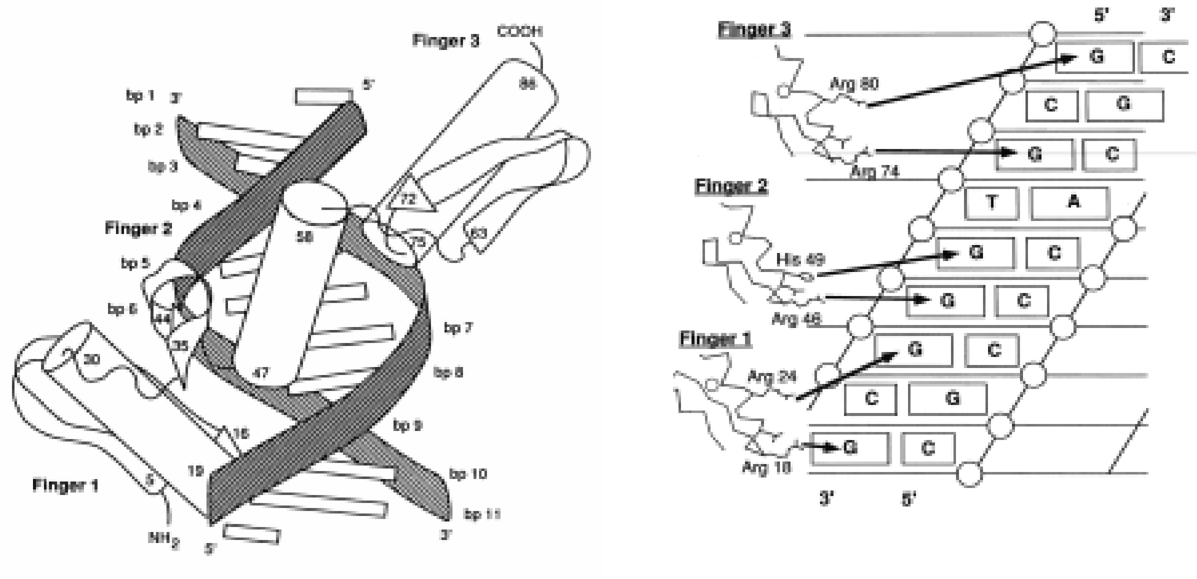

fig. 5 **The crystal structure of the three-module Zn-finger Zif268 complexed with cognate DNA** [41] shows that Zn-finger modules bind to B-DNA by inserting their α-helices into the major groove (left figure). Each Zn-finger module forms base-specific contacts with a three-basepair block (right figure).

The almost identical positioning of Zn-fingers in the major groove of basically undistorted B-DNA allows one to derive a rough stereochemical recognition code establishing general rules for which specific DNA sequence is recognized by a given set of amino acids on the basepair-facing side of the α-helices [35]. A prerequisite for the validity of this recognition code is that the Zn-finger modules adopt the aforementioned canonical binding geometry. However, this assumption does not hold true for each module in multi-modular Zn-finger proteins. For example, the 9-module transcription factor TFIIIA utilizes modules 1 –3 and 7 – 9 to bind DNA in a manner very similar to Zif268. But modules 4 – 6 function as connectors rather than DNA binding motifs, bridging the flanking 3-module clusters that do bind DNA [42,43].

Taken together, Zn-finger proteins bind cognate DNA by lining the major groove with the α-helices of two or more consecutive ββα-modules. Zn-finger domains show the most regular DNA binding geometry rendering them a suitable prototype for deducing a recognition code. However, non-canonical binding geometries limit its usefulness.

Leucine zipper DNA binding proteins

Leucine zippers are often involved in protein dimerization. They form heptad repeats in which every first and fourth position contains a hydrophobic residue with leucine predominantly occupying position four. During folding, which often goes along with dimerization, they build an α-helical coiled-coil in which leucine and other hydrophobic side chains of one α-helix interdigitate with leucine and hydrophobic side chains of the second α-helix like knobs into holes. This interleaved arrangement of leucines is reminiscent of a zipper, and thus designated leucine zipper. Since the seven residues in a regular α-helix turn for only 700 degrees rather than for two full rotations (720 degrees), two zipped α-helices are tilted against each other.

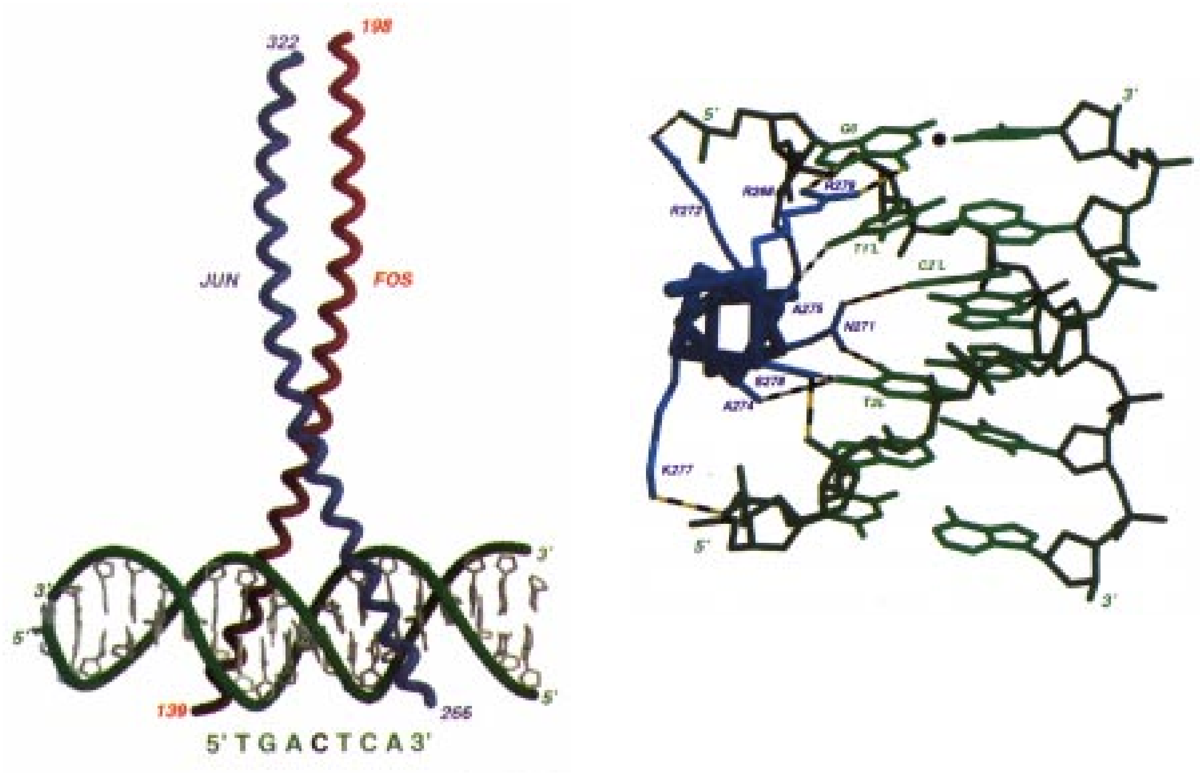

fig. 6 **The crystal structure of the Jun/Fos heterodimeric leucine zipper complexed with cognate DNA** [44] shows that two basic α-helical extensions bind into adjacent major grooves from opposite sides, thereby clamping the DNA between them (left figure). The basic helix of Jun forms a network of DNA contacts with the bases in the major groove (right figure).

DNA binding leucine zipper proteins are composite of the hydrophobic coiled-coil and an α-helical extension that is rich in positively charged residues. The tilt of the coiled-coil positions the positively charged extensions into two adjacent major grooves of cognate DNA. The crystal structure of the Jun/Fos heterodimeric leucine zipper complexed with cognate DNA [44] shows that the DNA is clamped between these two basic α-helical extensions forming base-specific contacts in the major groove (fig. 6). Both homodimeric, such as GCN4, and heterodimeric leucine zippers exist. The latter allow for a more complex combinatorial selection of DNA recognition sites. Similar to HTH and Zn-finger domains, leucine zipper DNA binding domains utilize the steric complementarity between α-helix and major groove to mediate base-specific DNA contacts.

β-sheet DNA binding proteins

β-sheet DNA binding proteins recognize cognate DNA either through β-sheet contacts in the minor groove, as observed for the eukaryotic TATA-box binding protein (TBP), or through β-sheet contacts in major groove. The latter class contains two homologous and structurally well-characterized members, bacterial MetJ repressor (MetJ-R) and phage P22 Arc repressor (Arc-R). The 53-residue Arc-R monomer consists of a ribbon-helix-helix (βαα) motif that requires dimerization for stable folding. Two βαα monomers intertwine in such a way that the two β-strands form an intermolecular antiparallel β-sheet. Two of these homodimers bind symmetrically to one *arc* operator, each by inserting its antiparallel β-sheet into the major groove of an operator half-site. The crystal structure of the Arc-R/operator complex [45] shows that the base-specific contacts are mediated by the antiparallel β-sheet in the major groove (fig. 7). Six polar side chains, including R, Q and N residues, on the major groove-facing side of the antiparallel β-sheet form an intricate network of hydrogen bonds with the bases and among each other. Additional phosphate backbone contacts are formed by the main chain H_N's of the second helix at the edge of the major groove. The structure of the Arc-R/operator complex demonstrates that the major groove of B-DNA is also well-sized to harbor a double-stranded β-sheet for DNA recognition, although α-helices are by far more often used for this purpose.

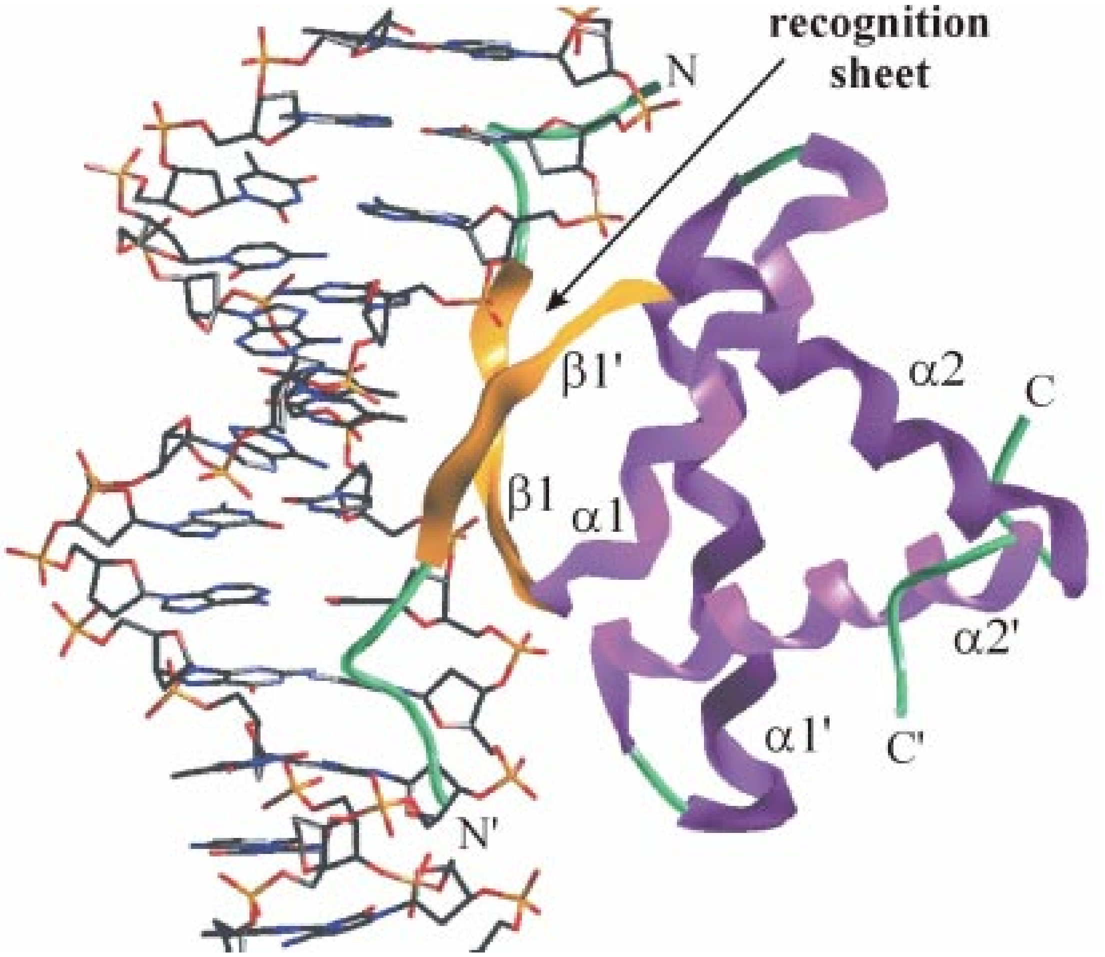

fig. 7 **The crystal structure of the Arc-R/half-operator complex** [45] shows DNA recognition by an antiparallel β-sheet in the major groove. The antiparallel β-sheet (yellow) of the Arc-R homodimer penetrates the major groove, thereby bringing the residues on its DNA facing side in contact with the bases of the operator half-site.

Immunoglobulin-like DNA binding domains

Transcription factors of the Rel family consists of two immunoglobulin-like domains, which are connected by a few flexible residues. They form homodimers [46,47] or heterodimers [48] through a β-sheet sandwich between their C-terminal domains. Unlike the DNA binding domains discussed before, which utilize α-helices or β-sheets for DNA recognition, Rel family dimers use loops from both their N- and C-terminal domains to mediate DNA contacts. The crystal structure of the NF-κB p50/p65 heterodimer (fig. 8), a classical member of the Rel family of transcription factors, shows that each Rel protein of the pseudo-symmetric heterodimer contacts the target DNA through five discontinuous loops wrapping around the virtually undistorted (14° bent) B-DNA helix [48]. The first loop in the N-terminal domain penetrates partially into the major groove to mediate base-specific contacts by four amino acid side chains. Except for one additional base-specific contact in the third loop, all other residues (11 residues in p50 and 8 in p65) form ribose and phosphate backbone contacts. Upon binding to the immunoglobulin(Ig)κB element, the p50/p65 heterodimer buries 3754 Å^2

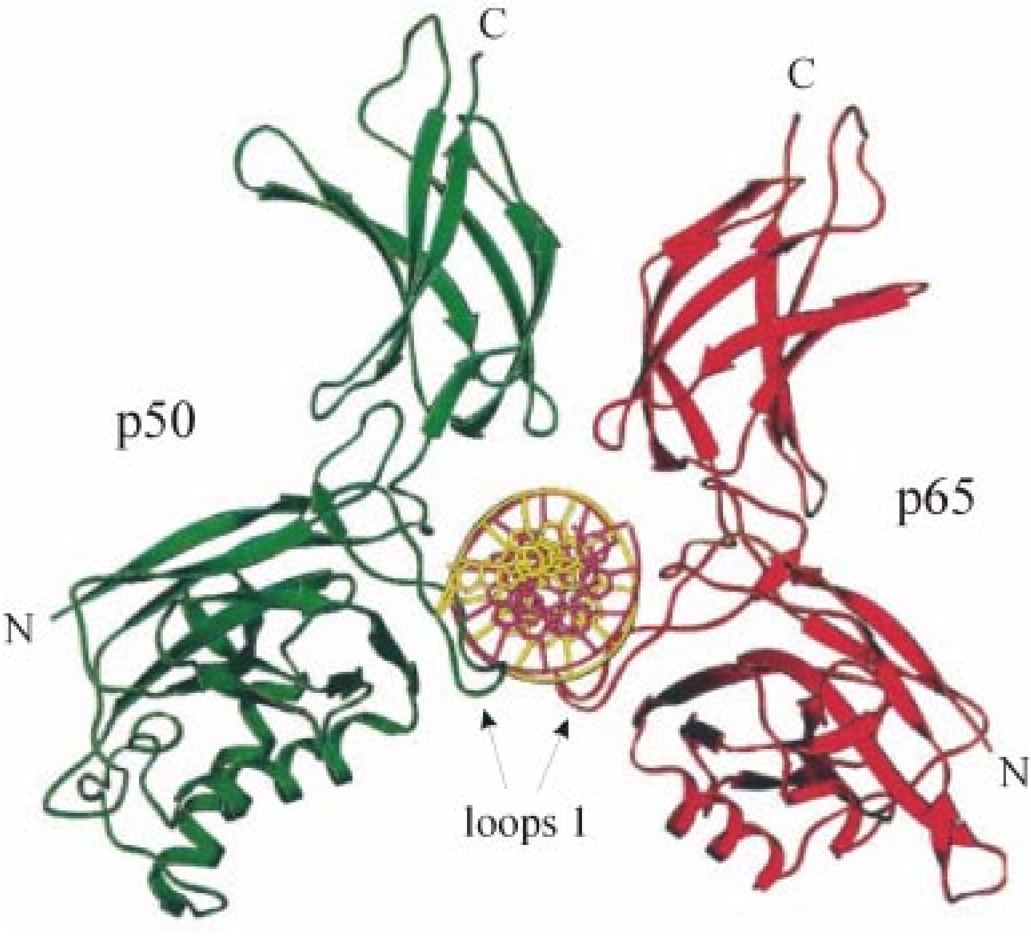

fig. 8 **The crystal structure of the NF-κB p50/p65 heterodimer**, an archetype for the Rel family of transcription factors, complexed with immunoglobulin(Ig)κB target DNA demonstrates DNA recognition exclusively through loops [48]. Each subunit of the pseudo-symmetric p50/p65 heterodimer contains two immunoglobulin-like domains connected by a 10-residue flexible linker. Loops from both the N- and C-terminal immunoglobulin-like domains contact DNA with loop 1 of the N-terminal domains almost penetrating the major groove of target DNA.

solvent-accessible surface area. This uncommonly large area [49] reflects the fact that the p50/p65 heterodimer embraces cognate DNA from all sides rather than docking from one side only.

In this DNA recognition mode, the use of flexible loops for forming DNA contacts may be favorable to fine-tune the binding geometry in the central tunnel accommodating the DNA. In contrast, DNA binding domains docking to the DNA from a single side are less constraint in their overall DNA binding geometry, and thus prefer rigid α-helices or β-sheets rather than flexible loops to mediate DNA contacts. These observations indicate that protein/DNA recognition requires a fine-tuned balance between backbone rigidity and local side chain flexibility to optimize the local positions of individual DNA contacts. An analogous balance has been observed for enzyme/substrate interactions for which the effects of backbone rigidity and side chain flexibility can be delineated. The importance of flexibility can be monitored through the rate of catalysis as a function of temperature. The role of rigidity can be deduced from investigating homologous enzymes from thermophilic, mesophilic and psychrophilic species.

Combinatorial control of transcription

The regulation of eukaryotic gene transcription often requires the coordinated binding of multiple transcription factors to the promoter/enhancer region. Many of these factors are independently controlled by different signal-transduction cascades. The combinatorial assembly of these factors is governed by DNA binding-induced protein-protein interactions and protein-induced DNA bending provided that the target DNA delivers a suitable composition and spatial arrangement of cognate sites [50]. This building block approach to recognize composite promoter and operator sites allows for the regulation of diverse patterns of genes by a small number of transcription factors and ensures a high specificity of transcriptional control by cooperative binding.

The crystal structure of the quaternary complex between nuclear factor of activated T-cells (NFAT), Jun, Fos and the distal antigen-receptor response element from the interleukin-2 gene promoter (fig. 9) [51], shows that NFAT adds an additional regulatory element to the ternary Jun/Fos/DNA complex (fig. 6). NFAT binds cooperatively to its DNA target site adjacent to the AP-1 site recognized by the Jun/Fos heterodimer, thereby forming a continuous 15 basepair ARRE2 binding site. The cooperativity stems from protein-protein interactions between the Rel family member NFAT and the coiled coil of Jun and Fos. The interaction surface contains a small hydrophobic patch on both Jun and Fos, but is otherwise hydrophilic. Although the extended leucine zipper of Jun and Fos bends towards NFAT and the NFAT-bound DNA sub-site bends towards the Jun/Fos heterodimer, the interaction surface between NFAT and Jun/Fos is not fully complementary. This sub-optimal fit may enable combinatorial assembly with other transcription factors. The example of the NFAT/Jun/Fos/DNA complex demonstrates that transcription factors can assemble to higher order complexes, thereby bringing together their individual target specificities in a structure-controlled manner.

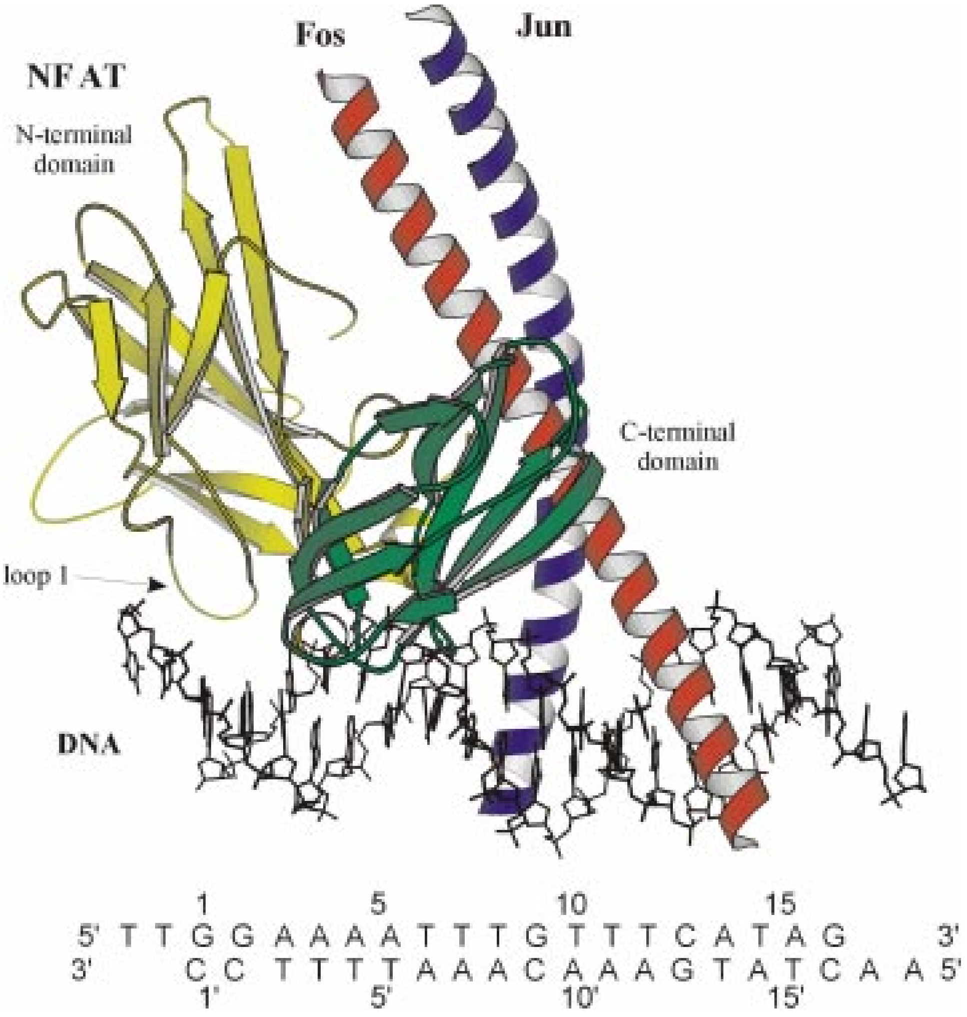
Fos
Jun
NF AT
N-terminal domain
C-terminal domain
loop 1
DNA
1 5 10 15
5' T T G G A A A A T T T G T T T C A T A G 3'
3' C C T T T T A A A C A A A G T A T C A A 5'
1' 5' 10' 15'

fig. 9 **The crystal structure of the DNA binding domains of NFAT, Jun and Fos complexed with cognate DNA** shows how individual transcription factors can cooperatively assemble on target DNA to mediate combinatorial regulation of transcription [51]. The DNA sequence of the distal antigen-receptor response element from the interleukin-2 promoter (ARRE2) is shown below the basepairs. The immunoglobulin-like factor NFAT is composite of an N-terminal (yellow) and a C-terminal (green) domain which both interact with the Jun/Fos zipper. DNA recognition through NFAT is similar to NF-κB with loop 1 contacting bases in the major groove.

DNA bending and DNA kinking proteins

A number of DNA binding proteins distort the conformation of canonical B-DNA. Except for the left-handed Z-DNA binding domain Zα, all of these proteins bend or kink DNA while retaining a right-handed B-DNA conformation in the segments surrounding such kinks and bends. In the following paragraphs, the principle mechanisms of DNA bending by protein domains of various classes is discussed.

DNA bending through backbone contacts: The CAP domain

Catabolite gene activator protein (CAP) activates transcription at more than 20 different promoters in *E.coli* when complexed with its allosteric activator cAMP. The crystal structure of CAP complexed with 30-base pair target DNA (fig. 10) shows that the symmetric CAP homodimer introduces an overall DNA bend of ~ 90° [37]. The bend is almost entirely achieved by two 40° kinks each between two TG/CA base pairs at position 5 and 6 on each side of the dyad axis of the complex. The C-terminal DNA binding domain of CAP has a (α+β)HTH fold with an additional N-terminal dimerization helix. Per half-site CAP forms three base-specific interactions through two arginines and one glutamate emanating from the recognition helix, which penetrates the major groove of the B-form DNA. More importantly, hydrogen bonds and ionic interactions between 13 protein functional groups and 11 DNA phosphates are formed per half-site. This intensive network of favorable backbone interactions makes the DNA bend around the CAP homodimer with the rigid fold of CAP governing the shape of the bound DNA. The specific recognition of the CAP binding site probably results from both the direct hydrogen bonds between three side chains and three bases, and sequence-dependent bendability of the CAP target DNA. The capability of CAP to severely bend cognate DNA is thought to play an important role in bringing CAP and upstream sequences into close contact with the RNA polymerase by folding remote upstream promoter regions back onto the transcription initiation region.

DNA bending by a β-stranded saddle: The TATA-box binding protein

The TATA-box binding protein (TBP) is the key component for the assembly of the multi-protein, pre-initiation complex responsible for transcription by all three RNA polymerases in eukaryotes. The phylogenetically conserved C-terminal domain of TBP binds to the TATA box present in almost all eukaryotic promoters. Data from recent genome

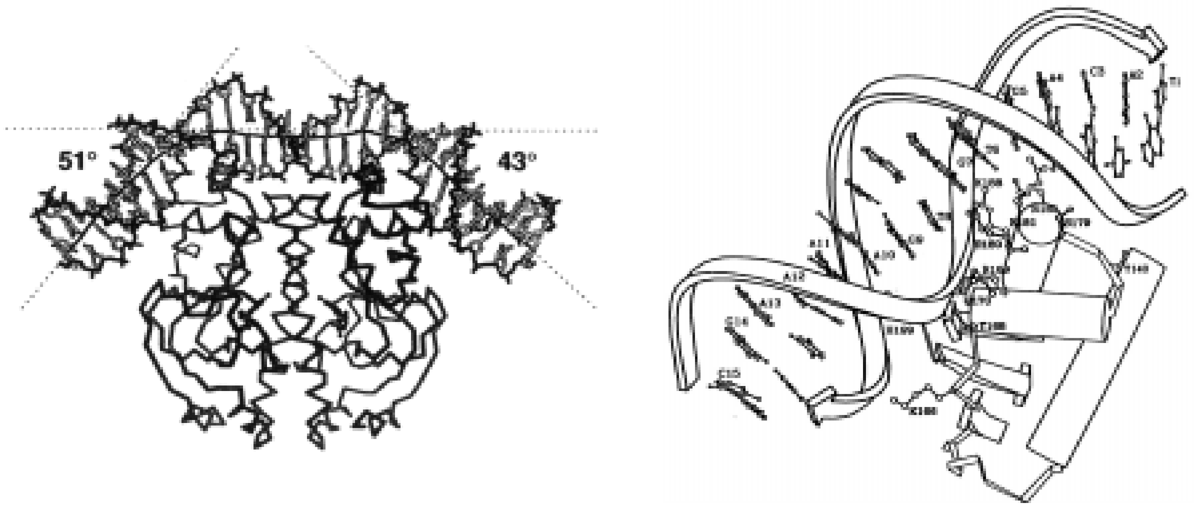

fig. 10 **The crystal structure of the CAP homodimer complexed with 30 base pairs of target DNA** shows how a (α+β)HTH-domain dimer can introduce two abrupt 40° kinks towards the major groove into bound DNA causing an overall DNA bend of ~ 90° [37] (left figure). The close-up of a CAP half-site shows that the recognition helix a3 penetrates the compressed major groove mediating three base-specific DNA contacts (right figure). However, the majority of protein/DNA contacts is targeted to DNA phosphates.

sequencing projects showed that the basal components of transcription complexes are homologous for archaea and eukaryotes including TBP, transcription factor IIB (TF-IIB) and the (A+T)-rich TATA element. In archaea, such as the hyperthermophile *P. woesei*, the assembly of TBP, TF-IIB homolog and RNA polymerase are sufficient to initiate promoter-specific transcription, while transcription initiation in eukaryotes requires additional factors. Thus, archaeal TBP, TF-IIB homolog and TATA-box homolog present a minimal intact biological system amenable to high resolution structural investigation of transcription initiation.

The crystal structure of the ternary complex between the TBP homodimer, TF-IIB and the TATA-box homolog from *P. woesei* (fig. 11) shows that the TBP homodimer forms a saddle-shaped concave surface with its 12-stranded antiparallel β-sheet [52]. This concave saddle exclusively binds to the minor groove of the 8-basepair TATA-box homolog by bending the DNA towards the major groove, thereby prying open the minor groove and underwinding the B-form DNA. Two sharp kinks are introduced into the DNA at both ends of the TATA-box by partial intercalation of two pairs of phenylalanines resulting in an overall bend of the double helix of 65°. DNA bending in the homologous eukaryotic ternary TBP/TF-IIB/DNA [53] and binary TBP/DNA [54,55] complexes is similar showing bends of 70° and 80°, respectively.

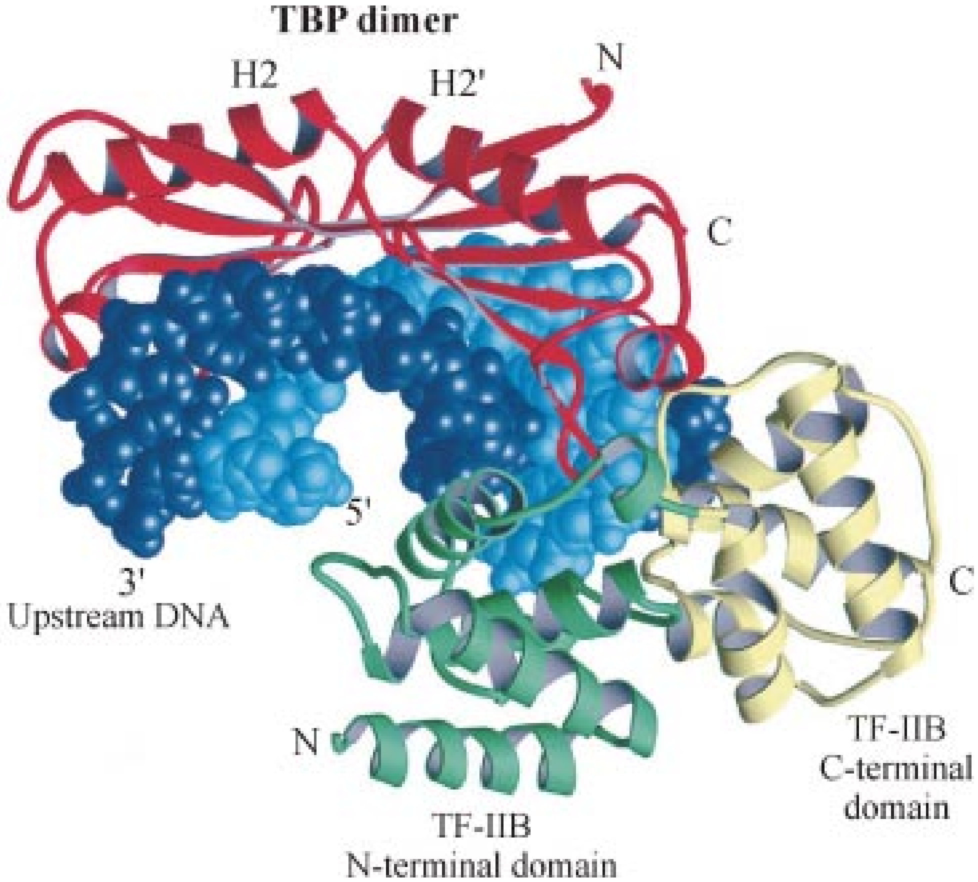

fig. 11 **The crystal structure of the archaeal ternary complex between the TBP homodimer, the transcription factor IIB (TF-IIB) homolog and the TATA-box homolog** shows how the β-stranded concave saddle of TBP (red) bends the DNA (blue) towards the major groove by phenylalanine intercalation at both ends and numerous interactions in the widened minor groove [52]. TF-IIB binds peripherally to TBP altering the geometry of the TBP/DNA complex only marginally.

Unlike the DNA double helix, TBP remains essentially unchanged when bound to the TATA-box homolog, as compared with the crystal structure of free TBP from *P. woesei* [56]. Side-chain/base interactions between TBP and the TATA element are primarily hydrophobic with 12 residues forming van der Waals contacts and only four residues forming hydrogen bonds. The DNA ribose-phosphate backbone is contacted by another 18 residues mediating 11 van der Waals and 7 hydrogen bond or salt bridge contacts.

In eukaryotes, TF-IIB mediates the contact between TBP and the RNA polymerase through interaction with the transcription factor TF-IIF [52]. TF-IIB, which binds to the C-terminal stir-up of TBP, exclusively contacts the DNA backbone with a total of 18 residues forming DNA contacts upstream and downstream from TBP. Comparison with binary TBP/TATA-element complexes [54,55] demonstrates that TF-IIB does not markedly modify DNA bending by TBP suggesting that TBP bends the TATA element in the transcription initiation complex of its own. Taken together, TBP achieves dramatic DNA bending by phenylalanine intercalation, a large number of hydrophobic contacts in the widened minor groove and an approximately equal number of contacts to the DNA backbone. These interactions make the DNA double helix adapt to the preshaped interaction surface of TBP, which remains its conformation.

DNA bending by an α-helical saddle: The SRY HMG domain

High mobility group (HMG) domains bind DNA in the minor groove, bend the DNA double helix and recognize four-way junctions and other irregular DNA structures [57]. The HMG family of eukaryotic proteins can be functional divided into sequence specific binding transcription factors, such as the sex-determining region Y (SRY) protein and lymphoid enhancer-binding factor (LEF-1 [58]), and non-sequence specific binding HMG domains, such as the upstream binding factor.

The solution structure of the SRY domain bound to cognate DNA (fig. 12) shows that the SRY domain bends the DNA by 70° - 80° and causes helical unwinding [59]. The SRY domain has the shape of a twisted L. The minor groove of the DNA binds in the concave surface of the L-shaped SRY domain through contacts with helix 1 and 3. The conformational change of the DNA is reminiscent of a classical induced-fit interaction in which the protein maintains its conformation. DNA bending occurs abruptly between basepair 5 and 6 by partial intercalation of an isoleucine, and between basepair 2 and 3 as a result of close contact between these basepairs and a ridge formed by a tyrosine and a lysine. Five residues of SRY widen the minor groove by forming a T-shaped hydrophobic wedge, which is in intimate contact with the bases. 11 residues at the surrounding wings of this wedge, including five arginines and two lysines, bind to the DNA sugar-phosphate backbone, thereby prying open the minor groove. This leads to a compaction of the major groove and strong bending of the DNA double helix. The almost 90° bend, commonly observed when HMG transcription factors bind to DNA, is thought to play an architectural role bringing previously remote components of the transcription activation complex into close proximity, thereby enabling them to interact.

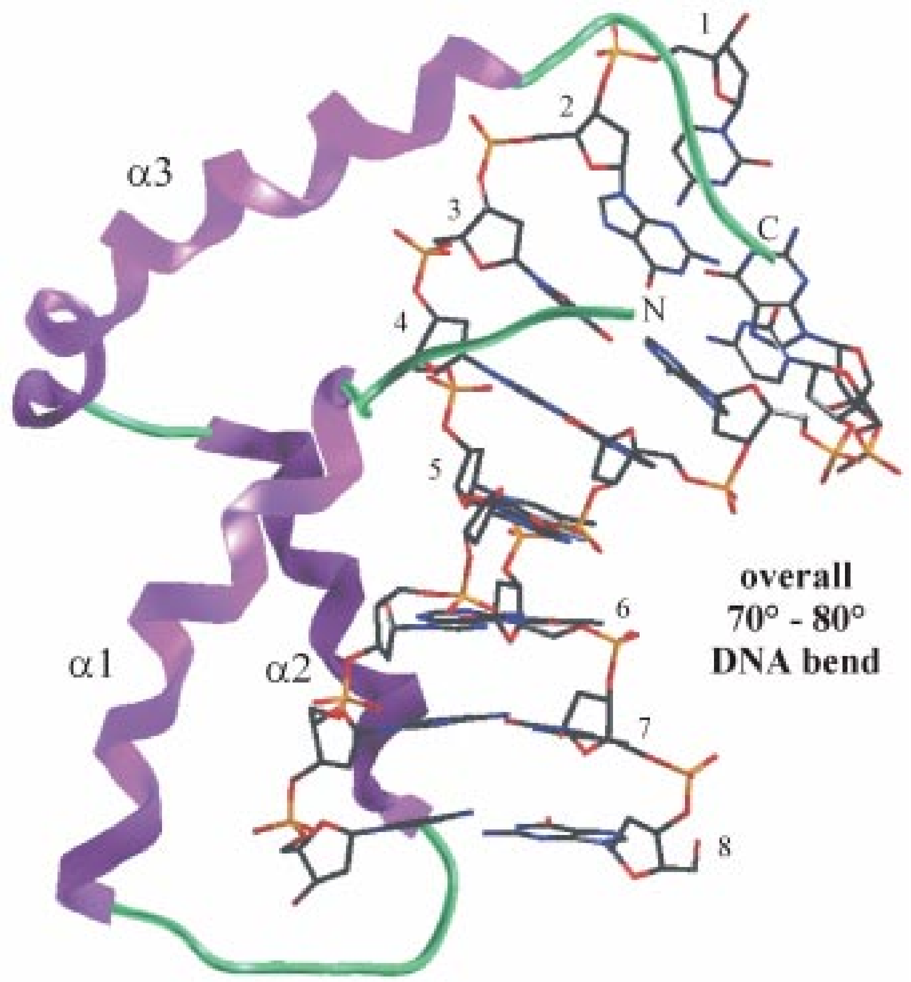

fig. 12 **The solution structure of SRY bound to target DNA** shows how the minor groove of B-DNA is pried open when the SRY domain binds resulting in a 70° - 80° bend of the DNA double helix [59]. SRY, which possesses a twisted L-shape, binds to the minor groove through helices α1 and α3, thereby forcing the DNA to adapt to its concave L-shaped interaction surface.

Summary

The information contained in the genome is read out through specific protein/DNA interactions. The major groove of B-DNA is the most favored docking site for proteins because it is easily accessible and information-rich, allowing bound protein to discriminate between the four bases by specific hydrogen bonds [60]. The diameter of an α-helix is complementary to the trough of the major groove rendering it a widely used protein motif for specific DNA read out among transcription factors. Further strategies of DNA recognition exist including DNA binding through β-sheets and through loops. HTH, Zn-finger and β-sheet DNA binding domains commonly dock to DNA from one side [49], whereas multi-domain transcription factors of the immunoglobulin-like class embrace cognate DNA.

DNA bound to protein adopts regular as well as various distorted B-DNA type of conformations. Proteins that bend or kink DNA generally form an exceptionally large number of contacts to the DNA backbone, thereby delivering the energy for forcing the DNA into a distorted conformation. In several cases, DNA bending proteins bind to the minor groove by prying it open leading to an overall bend towards the major groove. DNA kinking is often caused by partial intercalation of amino acid side chains. The DNA surrounding such kinks adopts an almost regular B-DNA conformation. In conclusion, DNA binding proteins can cause dramatic changes in the conformation of bound DNA, whereas the conformation of bound protein remains unchanged in most cases.

The left-handed Z-DNA conformation

DNA encodes information in two forms [61]. First, the linear sequence of four bases specifies the composition of proteins. Second, the three-dimensional structure determines the macromolecules with which DNA can interact. Structural information is involved in many interactions, such as stabilizing, replicating, reading, regulating, modifying or degrading DNA. A number of three-dimensional structures of DNA in complex with binding partners reveal that the DNA is amazingly flexible allowing bending [37,55,59] and other alterations to its shape [62]. Thus DNA can form a variety of different shapes [63] whose biological roles have been an intensively studied field.

One of the most dramatic conformational changes occurs when the right-handed B-DNA flips into the left-handed Z-DNA shape. The fundamental question of how nature uses the distinct conformation of Z-DNA, is the subject of the following paragraphs.

The structure of Z-DNA

First indications for a left-handed DNA helix were found more than 25 years ago. Circular dichroism (CD) spectroscopy of oligonucleotides with alternating deoxycytidine (dC) and deoxyguanine (dG) residues detected a characteristically inverted CD spectrum, as compared with B-DNA [64]. The physical reason for this finding remained a mystery until the high resolution crystal structure of the Z-DNA forming oligonucleotide d(CG)$_3$ revealed a left-

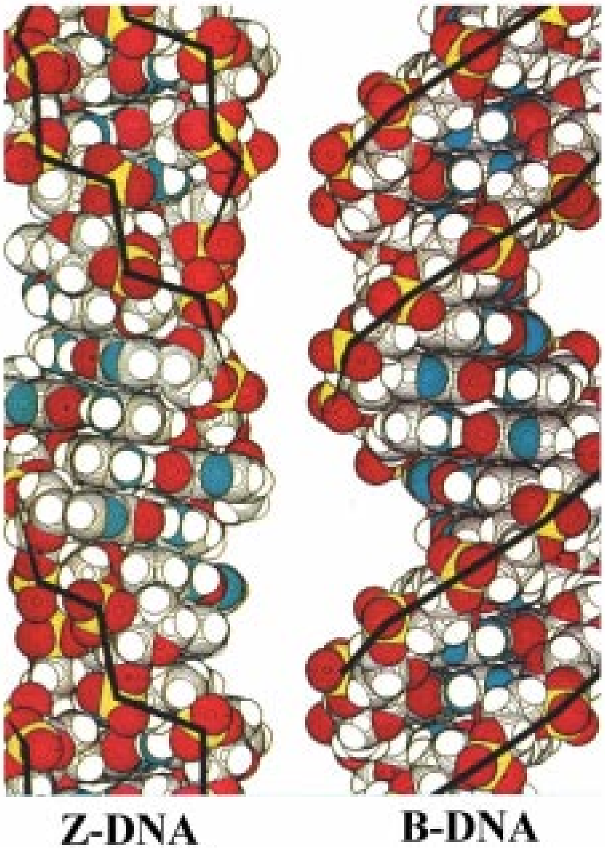

fig. 13 **Comparison between the structure of Z- and B-DNA.** The phosphate backbone is marked by black lines showing a characteristic zigzag line for Z-DNA. Z-DNA lacks a major groove, which is the principal binding pocket in most protein/B-DNA interactions. In B-DNA, the negatively charged phosphate backbone wraps around the core of the helix with the hydrophobic bases stacking in the center of the helix. In contrast, both hydrophilic backbone and hydrophobic bases are exposed in turns on the surface of Z-DNA offering both a polyanionic and a hydrophobic landing pad for interacting molecules.

handed double helix, which maintained Watson-Crick base pairing [65]. Sequences of alternating pyrimidine/purine nucleotides facilitate Z-DNA formation, with alternating dC/dG performing best [66-68].

The dramatic structural differences of Z-DNA compared to B-DNA are most strikingly marked by its zigzag shaped phosphate backbone (fig. 13) which gave rise to its name Z-DNA (‘Z’ abbreviates zigzag). In contrast to B-DNA, the repeating unit of Z-DNA is one d(CG) dinucleotide which is used to form the helix by continuous translation/rotation motions [69]. One helical pitch consists of 6 such pyrimidine/purine repeating units and spans 45 Å. Compared to B-DNA, which is made up of 10-basepair repeating units and has a helical pitch of 34 Å, Z-DNA is slimmer and stiffer. The left-handed helix exhibits a smaller van der Waals diameter of 18.2 Å than B-DNA with 19.3 Å. Unlike B-DNA, purine nucleotides adopt a *syn* conformation in Z-DNA. In response to the steric constrains exerted by the *syn* orientation, the sugar pucker of purines transforms into a $C_{3'}$-endo conformation, while the conformation of the glycosyl bondage of the pyrimidine nucleotides maintains the *anti* position during the transition to Z-DNA. As a result, the major groove, present in B-DNA, converts into a convex surface, whereas the minor groove deepens and narrows in Z-DNA (fig. 14).

On the convex surface of Z-DNA, the functional groups of the bases are especially exposed to the solution. Electrophilic chemicals such as diethylpyrocarbonate preferentially react with these more accessible bases of Z-DNA, enabling the mapping of Z-DNA segments within regular B-DNA sequences [70].

The transition from B → Z and vice versa is proposed to take place by a mechanism in which the Watson-Crick base-pairing is retained and one basepair at a time rotates 180^o into the inversely handed conformation. This B → Z flip migrates cooperatively along the double helix.

Since the Z-DNA conformation occupies an energetically less stable mode than B-DNA

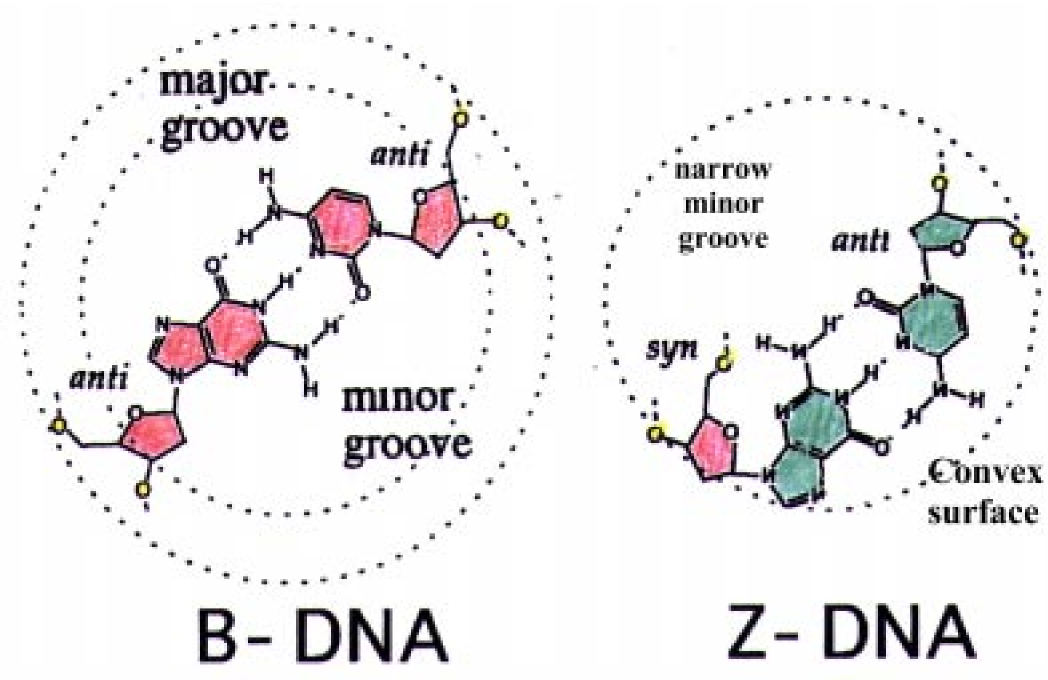

fig. 14 **Basepair conformation in B-DNA versus Z-DNA.** In B-DNA, all basepairs adopt an *anti* conformation with a $C_{2'}$-endo sugar pucker. By the contrary, purine basepairs of Z-DNA adopt a *syn* conformation with a $C_{3'}$-endo sugar pucker. The dotted circles illustrate that Z-DNA possesses a smaller diameter than B-DNA.

under physiological conditions, certain stabilizing effects must act to flip DNA into its Z-form. High concentrations of mono- or divalent cations effectively stabilize Z-DNA. Mg^{2+}, cobalt hexamine (II) and spermine, a naturally occurring polycation, stabilize Z-DNA by specific contacts [64,71,72]. Alternatively, Z-DNA can be stabilized by chemical DNA modifications. Methylation at position 5 of deoxycytosine as well as bromination favors the Z-DNA conformation [73,74], since these bulky hydrophobic side groups are less solvent-exposed in the Z-form. Furthermore, aqueous solutions of organic solvents, such as methanol or ethanol, in which the physicochemical activity of water is reduced, stabilize Z-DNA [75]. Although these modifications are easily accessible and offer straightforward avenues to study Z-DNA in vitro, they are non-physiological, meaning that Z-DNA formation *in vivo* must be caused by different mechanisms.

Z-DNA in the realm of biology

DNA naturally occurs in different topological forms. DNA topology is the sum of two contributions: firstly, the number of revolutions a polynucleotide strand wraps around the double helical axis is denoted twisting number or simply twist. Secondly, the number of turns the double helix itself makes about a virtual superhelical axis is designated writhing number or simply writhe. The writhe is often referred to as supercoiling. These contributions are not physically separated. This means that reverse sense (negative) supercoiling counteracts right-handed (positive) helical twist. Hence negative supercoiling can be used to wind the DNA double helix in the opposite, left-handed sense. In living organisms, DNA topology undergoes dynamic changes when nuclear processes, such as replication or transcription, are in work. *In vivo* two counteracting enzymes, topoisomerase I and II, control the supercoiling status of DNA.

The finding that Z-DNA forms under conditions of negative supercoiling *in vivo* brought the left-handed DNA conformation in the realm of biology [61] and offered novel avenues to study Z-DNA under physiological conditions [66,76,77]. Negative supercoiling as well as Z-DNA are states of elevated energy. The insertion of negative superhelical twists can provide the energy necessary to flip B-DNA into Z-DNA. The energy required to form Z-DNA can be determined by measuring the superhelical density of simple model systems, such as bacterial plasmids. The energy was found to be proportional to the square of the number of negative supercoils [78-81]. In this understanding, Z-DNA formation does no longer depend on specific sequences but is triggered whenever sufficient energy in form of negative supercoiling is reached. However, alternating pyrimidine/purine sequences, such as $d(CG)_n$ followed by $d(TG)_n$, $d(GGGC)_n$ and $d(TA)_n$, require lower energy levels to flip into the Z-conformation [80,82,83].

Energy calculations revealed that the formation of a B/Z-DNA junction has a free energy of $\Delta G = 4$ kcal/mol and thus constitutes a significant barrier for the formation of Z-DNA [78]. These biophysical data allowed the development of computer models to calculate the Z-DNA-forming potential of naturally occurring sequences [82,84]. Investigation of 137 fully sequenced human genes revealed that 98 genes contain potential Z-DNA-forming sequences. Interestingly, these sequences were detected 10 times more frequently in the 5’ upstream than 3’ downstream regions of these genes [82,84]. This non-random distribution suggests that Z-DNA formation corresponds to transcription processes which are initiated in the 5’ upstream region of genes. The demonstration that a moving RNA polymerase generates negative superhelical stress at its 5’ side as it ploughs through the DNA double helix [85], supports this notion. The RNA polymerase thus provides a plausible causative mechanism for initiating Z-DNA formation *in vivo.*

Z-DNA in prokaryotes

The existence of Z-DNA in bacteria was investigated by three different indirect approaches. Firstly, d(CG)$_n$ inserts within bacterial plasmids were shown to form Z-DNA *in vivo* by modification with electrophilic chemicals (see paragraph „The Z-DNA structure“), such as osmium tetroxide, potassium permanganate and others [86,87]. Secondly, these results could be confirmed by UV cross-linking that allows one to measure the unrestrained superhelical density of such plasmids when the bacteria have been previously treated with topoisomerase inhibitor [88]. Thirdly, Z-DNA segments are not methylated within *E.coli* cells. A construct embedding a *Eco*RI site in a Z-DNA-forming sequence functions as a reporter for torsional strain within the bacteria because non-methylated *Eco*RI sites are susceptible to restriction digest. By assaying the methylation of the *Eco*RI site, which can only occur when the insert has a B-DNA conformation, Z-DNA formation could be measured in *E.coli* without external perturbation. Activation of transcription as well as defects in topoisomerase I could be shown to enhance Z-DNA formation [89-91]. In contrast, *Enterobacter, Klebsiella* and *Morganella* did not form Z-DNA in this experiment [91].

Z-DNA in eukaryotes

Although direct Z-DNA assays are difficult in eukaryotes, a body of evidence points to Z-DNA formation in eukaryotes. For these more complex organisms, Z-DNA assays base on poly- and monoclonal antibodies [92]. The first antibodies were gathered from sera of humans who suffered from autoimmune diseases, especially lupus erythemantosis. These patients produced antibodies specific for Z-DNA during the exacerbations of the disease. Knowing that Z-DNA triggers immune response, antibodies were raised in rabbits and sheep. Antibody staining experiments with both fixed [93] and unfixed polytene chromosomes of *Drosophila melanogaster* showed intense staining in the interband regions but not in the bands of these chromosomes [94]. Strongest signals were found in the puffs which are known for high transcription activities [95].

Further evidence for a link between Z-DNA formation and transcription came from antibody staining experiments on ciliated protozoa. The micronucleus, which is responsible for genetic reproduction, was negative whereas the macronucleus, where transcription takes place, was stained [96].

In mammalian systems, evidence for the existence of Z-DNA *in vivo* was provided by studying metabolically active permeabilized nuclei. Nuclei were made by embedding intact cells into agarose and permeablizing their plasma membrane by careful treatment with detergent. In these experiments, Z-DNA was assayed by diffusing biotin-labeled anti-Z-DNA monoclonal antibodies into the nuclei and detected with radioactive labeled streptavidin in a subsequent step [97]. The results revealed that negative torsional strain affects the amount of Z-DNA, which increased remarkably when transcription had been activated, but was largely unaffected by replication [98].

The same research group developed a method for cross-linking anti-Z-DNA antibodies to DNA by exposing murine U937 cells to a 10 ns laser-pulse at 266 nm wavelength [99]. The chromosomal DNA was fragmented in situ by restriction digestion. Antibody cross-linked fragments were purified by streptavidin/biotin affinity chromatography and thereafter released by antibody proteolysis. Applying hybridization or PCR amplification techniques, three transcription-dependent Z-DNA forming segments in the 5' region of the human *c-myc* gene could be mapped [100]. For the human corticotropin hormone releasing gene, increasing and decreasing Z-DNA formation could be shown to correlate with gene up- and down-regulation, respectively.

The common denominator of these experiments suggests that Z-DNA forms largely in the 5' region of eukaryotic genes and that Z-DNA formation correlates with transciptional activity. These findings raise questions concerning the function of Z-DNA in such genes during transcription.

Potential roles of Z-DNA *in vivo*

In the simplest model Z-DNA may function by blocking interactions that would take place if its sequence is in the B-DNA conformation. For example, the RNA polymerase may not transcribe through Z-DNA segments, as was demonstrated for *E.coli* [101]. Thus, Z-DNA may ensure an adequate spacing between successive RNA-polymerases, since transcription cannot be initiated again before the negative superhelical stress in the promoter region is relaxed. This mechanism could prevent side reactions in highly induced genes, such as non-functional trans-splicing in eukaryotes.

Another model uses Z-DNA segments to buffer negative topological strain that arises when intact DNA duplexes are intertwined during recombination of homologous chromosomal domains [102]. For instance, the Z-DNA-forming sequence $d(CA/GT)_n$ promotes recombination in yeast [103]. However, it was shown to be less efficient than $d(CG)_n$ in human cells [104,105] demonstrating that recombinogenic sequences in humans are also potentially Z-DNA forming sequences.

Several reports of chromosomal breakpoints in human tumors contain potential Z-DNA-forming sequences, although no causal relationship has been demonstrated [106-110]. Finally, Z-DNA may be involved in the arrangement of nucleosomes and thereby affect the organization of chromosomal structure [111].

In the end, these hypotheses suggest a supporting rather than a leading role for Z-DNA in biology. Recent evidence suggests that Z-DNA may play an auxiliary role in mammalian RNA editing which modifies the genetic program of a cell. The following introduction describes an enzyme that combines two intriguing facets, RNA editing and Z-DNA binding activity, on one polypeptide chain.

RNA editing

RNA editing interrupts the linear flow of information from DNA to proteins. It describes the alteration of RNA information other than by RNA processing, such as splicing, capping, polyadenylation and the creation of hypermodified bases. RNA editing brings in diversity at a novel step: after transcription and in most cases before RNA processing. RNA editing is widespread among higher eukaryotes, reaching from fungi and plants to mammals and is even found in clinically relevant viruses. In humans it plays a role in physiology as well as pathogenesis. Since RNA editing is a „leaky" process it allows smooth evolution [112,113]. Mutated proteins can be tested while their non-mutated predecessors are still present. Since higher organisms do not evolve by means of high mutational rates, as it is common to prokaryotes, they may compensate by a sophisticated regulation of modification mechanisms, such as RNA editing. In this context, the lack of RNA editing in prokaryotes, as suggested by current knowledge, may be a consequence of evolutionary competition, in which complex modifications at the RNA level have lost or have proved unnecessary in prokaryotes.

From a mechanistic point of view, RNA editing can be divided into insertion/deletion editing, in which RNA is cleaved and bases are added or removed, and substitution editing, in which single bases are changed by deamination.

Insertion/deletion RNA editing

The first and most extreme example of RNA editing was discovered in mitochondrion-encoded mRNA of the kinetoplastid protozoan, *trypanosoma brucei*, 14 years ago [114]. In this trypanosome up to hundreds of uridine residues (U) are inserted or deleted in a single mRNA. Subsequent investigations determined that many mitochondrion-encoded RNAs are edited in these organisms [115,116]. Further insertion/deletion editing processes were found in the kinetoplastids (U insertions/deletions) *Leishmania tarentolae* and *Crithidia fasciculata* and in the slime mold *Physarum polycephalum* (U, G, A, C insertions).

Most knowledge about insertion/deletion editing has been accumulated for kinetoplastids. It generates initiation and termination codons and corrects internal frameshifts [117]. It was shown to take place after transcription [118], and some indications suggest that it precedes RNA processing and polyadenylation [119,120]. However, in some cases transcripts are already polyadenylated while they are still partially edited. This demonstrates that the timing of the editing process is somewhat flexible with regard to RNA processing.

Recently, the catalytic mechanism of insertion/deletion editing was deciphered [121]. Small trans-acting guide RNAs (gRNAs), 60-80 bases in length, supply the genetic information necessary to guide this type of editing (fig. 15). Their 5' termini form Watson-Crick base pairing with the pre-edited mRNA, whereas their 3' ends pair with the edited product, leaving an unpaired region in their center. In a series of enzymatic steps, a pre-edited mRNA is cleaved within its unpaired region. For insertion editing, free UTP is added to the 3' cleavage site, while for deletion editing, unpaired uridine residues are hydrolyzed from the 3' cleavage site. The number of both inserted and deleted uridine residues is directed by base-pairing with the gRNA. Finally, the separated mRNA strands are religated.

Viral examples of insertion editing are found in the P gene RNA of *Parmyxoviruses* [122] and in the RNA of a glycoprotein of the *Ebola virus* (for review see [123]).

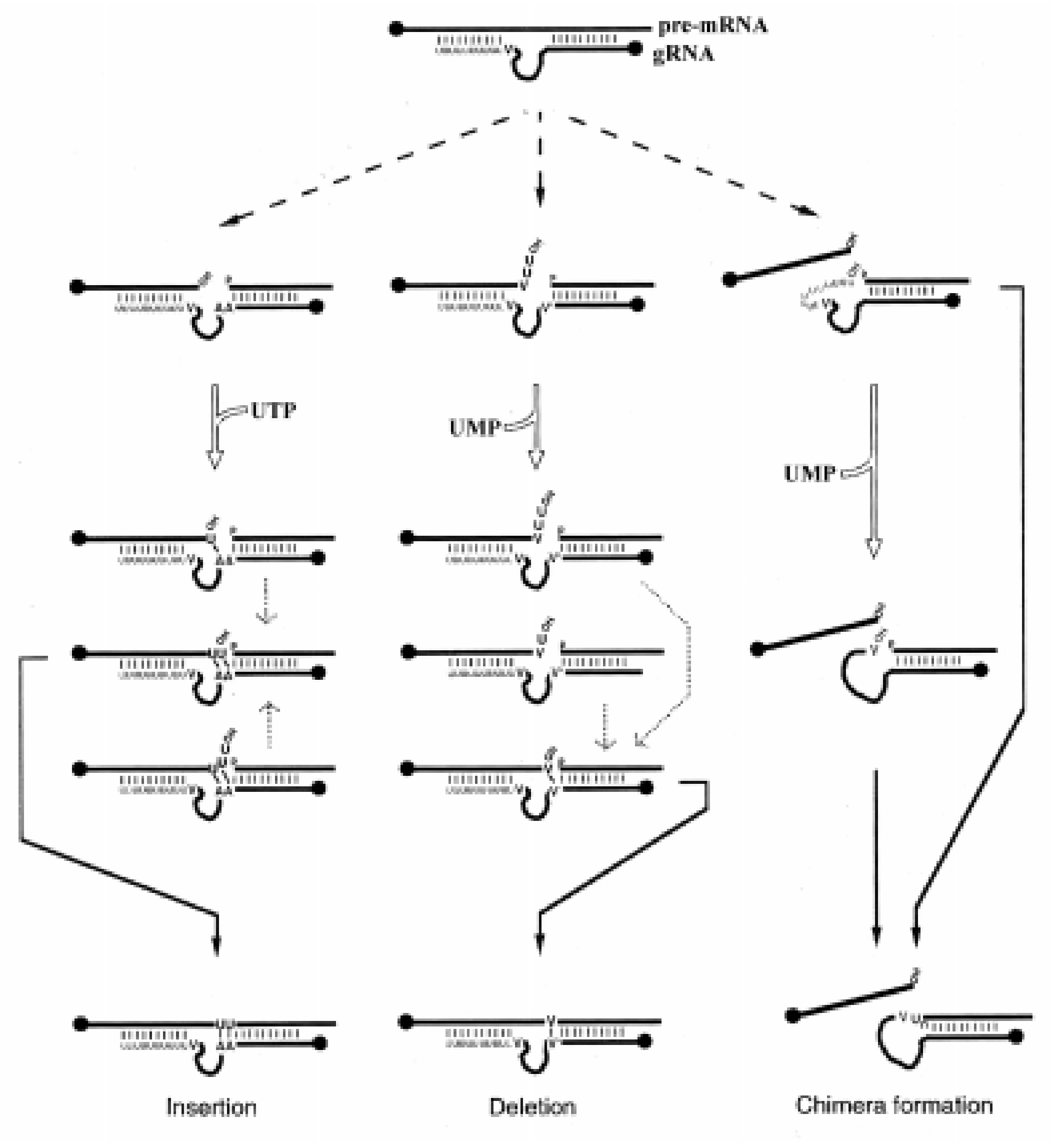

fig. 15 **Insertion/deletion RNA editing in trypanosomes** (adapted from [121]). A specific guide RNA (gRNA) hybridizes with the single-stranded pre-mRNA template. The pre-mRNA strand is enzymatically cleaved in the center, where a certain number of A nucleotides in the bulge loop of the gRNA governs the insertion of complementary U nucleotides into the mRNA strand. In U deletion editing, the lack of complementary A nucleotides in the bulge loop of the gRNA leaves a certain number of U nucleotides in the mRNA strand unpaired, which are therefore amenable to enzymatic hydrolyzes.

Substitution RNA editing

Substitution editing encompasses a series of distinct and probably separately derived traits. This RNA modification process is common for higher eukaryotes and is also involved in human disease, as shown for the mRNA of Wilms tumor-1 [124] and neurofibromatosis type 1 [125]. The examples of substitution editing, which have been discovered so far (table 1), can be divided into two groups:

1) cytosine to uridine (C to U) conversions and vice versa
2) adenine to inosine (A to I) conversions

table 1 **Substitution RNA editing** (adapted from [126] and updated).

Organism	Genome	RNA	Substitution
Physarum polycephalum [1]	Mitochondria	mRNA	C to U
Acanthamoeba castellani	Mitochondria	tRNA	U to A, U to G, A to G
Spizellomyces punctatus	Mitochondria	tRNA	U to A, U to G, A to G
Vascular plants	Mitochondria, chloroplasts	mRNA, tRNA, rRNA	C to U
Vascular plants	Mitochondria	mRNA, tRNA, rRNA	U to C
Drosophila melanogaster	Nucleus	Na^+ channel mRNA, 4f-rnp mRNA	A to I
Squid	Nucleus	Kv2 K^+ channel mRNA	A to I
Marsupials	Mitochondria	tRNA	C to U
Mammals	Nucleus	apoB [2] mRNA	C to U
Mammals	Nucleus	WT1 [3] mRNA	U to C
Mammals	Nucleus	AMPA, Kainate and 5-HT_{2C} receptor-mRNA, ADAR2 mRNA	A to I
Mammals	Nucleus	tRNA	C to U, U to C
HIV	Virus	TAR RNA	A to I
Hepatitis Delta Virus	Virus	RNA genome or RNA	U to C or A to I

[1] Also performs insertion/deletion RNA editing. [2] apolipoprotein B. [3] Wilms tumor susceptibility.

Cytidine to Uridine RNA editing (and vice versa)

The tissue-specific editing of apolipoprotein B (apoB) mRNA in mammals is the best studied example for cytidine to uridine (C to U) RNA editing. This early posttranscriptional event converts a glutamine codon (CAA) into a stop codon (UAA) [127] resulting in the production of the 300 kDa smaller apolipoprotein B48 (apoB48). ApoB48 is required for dietary lipid absorption, whereas the full-length apoB100 functions in transporting endogenously synthesized cholesterol and triglyceride in the circulation. Furthermore, apoB100 is the key surface marker in low density lipoproteins (LDL). Elevated LDL levels lead to atherogenic disease including heart attacks which are the most frequent causes of death in modern societies [128,129].

In addition to this stop codon formation at nucleotide position C6666, an alternate C to U site editing event was discovered at position C6802 [130]. However, this threonine (ACA) to isoleucine (AUA) conversion occurs with 20-fold lower frequency. Moreover, this portion of the mRNA is lost due to the activation of a cryptic polyadenylation site on mRNAs edited at C6666, suggesting that this alternate editing site is of no biological consequence.

Using an in vitro system, it was demonstrated that C to U editing of apoB is catalyzed by a cytidine deaminase (designated APOBEC1 for apoB mRNA editing catalytic subunit 1), which shows homology to the *E. coli* cytidine deaminase [131,132]. The *E. coli* cytidine deaminase has two core domains of similar tertiary structure [133]. One contains the active site with zinc bound at its center. The catalytic domain of APOBEC1 possesses a tertiary structure and a mechanism of catalysis that is conserved among cytidine deaminases [134]. APOBEC1 forms a homodimer that binds specifically to the 22 bp sequence of apoB mRNA. It acts in a multiprotein complex (27 S; 1400 kDa), designated editosome. APOBEC1 is expressed in testis, ovary and spleen, which are tissues that do not produce apoB100, suggesting that the editing of apoB mRNA is very efficient in these tissues.

Less is known on a number of other examples of C to U editing. In Wilms tumor susceptibility (WT1), in which the codon of leucine 280 (CUC) is converted into a proline codon (CCC) [124], the editing event suppresses the inhibitory action on the early growth response 1 promoter and may be involved in development and tumorigenesis. In rats, the major cytosolic tRNA for aspartate undergoes both C to U and U to C editing of two nucleotides adjacent to the anticodon loop to generate the major tRNA species [135]. In the mitochondria of marsupials, the anticodon of glycine tRNA is edited in the C to U direction, which converts it to an aspartate tRNA [136]. In vascular plants, RNA from mitochondria and chloroplasts is extensively edited showing both C to U and U to C conversions [137-139]. RNA editing has not been shown for bryophytes (mosses and liverworts) and chlorophytes (green algae). Plant mRNA editing corrects multiple gene-encoded missense codons, which include translation initiation and stop codons, thereby allowing functional protein synthesis. Although these examples show a common editing reaction, their scheme of site recognition differs. However, a common scheme has been observed for editing of tRNAs and introns which is always carried out in base-paired stems, where mismatches in base-pairing mark editing sites [140].

Adenosine to Inosine RNA editing

The most important excitatory neuro-receptors in the mammalian brain, the glutamate-gated cation channels, are the best-studied systems for adenine to inosine (A to I) substitution editing. Glutamate mediates fast excitatory neurotransmission by activating cation-selective channels with distinct gating kinetics, ion permeabilities and pharmacological properties. Glutamate-gated receptors transmit the majority of rapid excitatory signals in the central nervous system [141]. They are obligate parts of the hardware that provides for synaptic plasticity, learning and memory. They also mediate the toxicity of excess glutamate release from cells in pathological conditions, such as stroke [142,143].

Glutamate receptors are divided into three classes according to their sensitivity to the glutamate-imitating agonists AMPA (α-amino-3-hydroxy-5-methylisoxazole-4-propionic acid), NMDA (N-methyl-D-aspartate) and kainate. Individual sensitivities are governed by the subunit composition of glutamate receptors (GluR). AMPA receptors consist of four different subunits, designated GluR-A, GluR-B, GluR-C and GluR-D, while kainate receptors can be assembled from three different subunits, GluR5, GluR6 and GluR7. Of a total of approximately 18 GluR transcripts, five undergo A to I editing at up to three different codons. These include the channel-forming subunits GluR-B, GluR5 and GluR6.

In 1991 GluR-B mRNA was found to undergo A to I editing which converts a glutamine (CAG) to an arginine (CGG) residue (abbreviated as ‘Q/R site editing’) in the ion channel-forming transmembrane domain II. The positively charged arginine residue causes a dramatically reduced calcium permeability of this channel. Heterozygous mice that harbor an editing-incompetent GluR-B allele and hence express unedited GluR-B subunits in principal neurons and interneurons, develop epilepsy-like seizures and die by an age of 3 weeks, showing that GluR-B mRNA editing is essential for brain function. Moreover, AMPA receptors undergo another arginine (AGA) to glycine (GGA) editing event (R/G site editing) [144]. Here the adenine to inosine editing sites are determined by an intron that is present in GluR-B, -C and -D but lacks in GluR-A. Edited channels exhibit significantly shorter relaxation times and may thus perform better in integrating incoming signals.

Kainate-sensitive glutamate receptors are subjected to Q/R site editing, in which subunits GluR5 and GluR6 are edited in rat brain with an efficiency of 40 % and 80 %, respectively. For the transmembrane domain I, encoded by subunit GluR6, two further editing sites have been described. These editing events convert an isoleucine (AUU) to a valine codon (GUU) and a tyrosine (UAC) to a cysteine codon (UGC), designated I/V and Y/C site editing, respectively. In adult rat brain, two thirds of all GluR6 subunits are edited at all three sites.

Recently, A to I editing was discovered in the mammalian serotonin-2C receptor (5-HT_{2C}) mRNA [2]. 5-HT_{2C} receptors, which are widely expressed in the central nervous system, regulate a broad spectrum of physiological effects, such as appetite, mood and consciousness, and are involved in diseases, such as eating disorders, depression and epilepsy [145]. RNA editing changes three amino acids in the second intracellular loop of the 5-HT_{2C} receptor resulting in a 10 – 15-fold reduced G-protein coupling efficacy. This leads to a decreased activation of phospholipase C, which transmits the serotonin signal to downstream effectors.

In *Drosophila melanogaster*, A to I editing occurs in the highly conserved regions of a Na^+ channel [146]. Editing of the Na^+ channel mRNA requires cis-acting elements that are thought to form double-stranded RNA secondary structures similar to the double-stranded RNA structures necessitated for editing of GluR transcripts. Another example of A to I editing is the *D. melanogaster* gene *4f-rnp* [147]. An astonishingly large number of adenosines are converted to inosines in the transcript of *4f-rnp* which encodes an RNA-binding protein. The attractiveness of investigating A to I editing in *D. melanogaster* stems from the opportunity to utilize powerful genetic experiments.

A to I editing of mRNA encoding squid Kv2 K^+ channels alters amino acids in the pore region and in the „voltage sensor“ region of the channel, thereby changing the rates of channel closure and slow inactivation [148]. Voltage-gated ion channels, such as Na^+ and K^+ channels, play a crucial role in the generation and propagation of action potentials in electrically excitable cells. K^+ channels are tetramers with each monomer comprising six transmembrane segments. Transcripts including segments 4 – 6 of the Kv2 K^+ channel were investigated for this study.

Intriguingly, ion channels appear to be a primary target for A to I editing in higher eukaryotes. This suggests that RNA editing delivers a successful avenue by which a channel can modulate its activity allowing the nervous system to smoothly adapt to different environmental conditions.

The RNA editing enzymes ADAR1/2

A to I RNA editing leads to amino acid substitutions because the translation apparatus (and also the reverse transcriptase of retroviruses) read inosine as guanine. In vitro studies of the editing of AMPA-sensitive glutamate receptors have established that the A to I conversion is a hydrolytic deamination [141,149,150]. This enzymatic reaction is most likely accomplished by two related double-stranded RNA (dsRNA) deaminases, ADAR1 and ADAR2.

ADAR1 and ADAR2 are ubiquitously expressed and located in the nucleus of most cells [151,152]. ADAR1, which is the best studied dsRNA deaminase, is found in all higher eukaryotes. Originally, ADAR1 was described as a dsRNA unwinding activity [153,154]. Further studies indicated that unwinding of RNA double helices is a secondary effect resulting from the conversion of A-U to I-U base pairs, which destabilize the RNA double helix [155]. Using homology cloning, a variant of ADAR2, designated dsRNA editase 2 (RED2), was recently discovered showing 60 % amino acid identity with ADAR2 [156]. However, the brain-specific RED2 lacks deaminase activity on all known A to I editing targets including synthetic and naturally isolated dsRNA substrates.

Both ADAR1 and ADAR2 specifically act on dsRNA, neither single-stranded RNA (ssRNA) nor DNA are edited [1,151]. dsRNA is formed by base-pairing between an exonic sequence around the to-be-edited adenosine and an intronic „editing complementary sequence“ (abbreviated ECS). Deletion of the ECS leads to a complete loss of editing indicating that the ECS is essential for A to I editing [157]. Furthermore, the insertion of the ECS of the Q/R site of GluR-B induced editing in the homologous exons of GluR-A, -C and -D, which naturally lack this intronic ECS and are thus not edited [158]. The requirement for intronic sequences raises the possibility that a vast, yet silent source of DNA information, the intron pool, contributes to the formation of diversity in nature [112].

Regarding the timing and localization of A to I editing, the requirement for intronic sequences unambiguously sets this process before RNA splicing and restricts it to the nucleus. Since splicing occurs immediately after or parallel to transcription, ADAR-mediated RNA editing must take place almost co-transcriptionally.

The spacing between the exonic editing site and the intronic ECS varies considerably ranging from 40 nucleotides for the R/G site of AMPA-transcripts [159] to approximately 2000 nucleotides in the case of the Q/R site of GluR5 and GluR6 transcripts [160]. The adenosine that is to be edited can be base-paired, mismatched or reside in a short loop, as predicted by RNA secondary structure programs [159]. Comparison of known dsRNA editing sites does not reveal consensus sequences suggesting that a complex dsRNA structure poses the recognition motif for dsRNA deaminases.

The substrate specificity of ADAR1/2

A body of data has been accumulated concerning substrate specificity of ADAR1 in vitro. dsRNAs of approximately 100 bp are edited most efficiently [161], whereas templates shorter than 36 bp are modified poorly [162]. ADAR1 has a 5’ neighbor preference for A and U, but no 3’ neighbor preference [162]. It selectively modifies the A’s near the strand termini of short dsRNAs. However, beyond these glimpses of specificity, both biochemically purified ADAR1 or in *Xenupus* oocyte extracts expressed ADAR1 edit promiscuously [163,164]. Although a number of approaches have been used to achieve specific editing in in-vitro systems, it has not been possible to demonstrate the selectivity observed *in vivo* [151]. The addition of nuclear extract to in-vitro systems was shown to boost site selectivity and editing efficacy of ADAR1 suggesting that auxiliary factors are necessitated for ADAR1-mediated editing.

In *in-vivo* systems, ADAR1 edits the R/G site of AMPA-receptor subunits and certain intronic adenosines, such as the hotspot1 in the GluR-B intron 11, with high efficiency, but shows low editing efficiencies on all other native GluR transcripts. In contrast, ADAR2 edits the Q/R and R/G sites in GluR-B with high efficiency and site selectivity without the addition of auxiliary RNA or protein factors [165]. Of the three codons (labeled A, C and D) altered by A to I editing in the serotonin-2C receptor, codons A and C, but not codon D, are edited by ADAR1 with high efficiency in cotransfected HEK293 cells [2]. ADAR2 edits codons C and D with high efficiency, but is less efficient than ADAR1 at codon A. These results demonstrate that ADAR1 and ADAR2 possess different substrate specificities which may overlap for some editing sites. This suggests that both enzymes need to work together for complete editing of various pre-mRNA substrates.

Recently, it has been shown that ADAR2 regulates its own alternative splicing by A to I RNA editing [166]. ADAR2 pre-mRNAs are spliced at either a proximal or distal 3' acceptor site with the genomic sequences AA and AG, respectively. Editing converts the AA of the proximal site into an AI (read as AG), which is the preferred consensus sequence of 3' acceptor sites. Using this alternative 3' acceptor site, 47 nucleotides are inserted into the ADAR2 transcript causing a frameshift. By using an alternative translation initiation site, the frameshifted transcript can be converted into active ADAR2 protein. The modulation of alternative ADAR2 splicing by ADAR-mediated editing may represent a negative autoregulatory mechanism by which ADAR2 can prevent its own overexpression and thus aberrant editing.

The domain structure of ADAR1/2

Sequence comparison of mammalian RNA editing enzymes shows that they share a homologous zinc-dependent catalytic deaminase domain (fig. 16). Interestingly, both the adenosine and the cytidine deaminase domains of ADAR1/2 and APOBEC1 (see „C to U editing"), respectively, show homology to the (mono-nucleoside) cytidine deaminase domain, but not to the adenosine deaminase domain, of *E.coli* [126]. In addition, the ADAR enzymes have two homologous dsRNA binding domains (DRBD) in common. Deletion analysis has shown that ADAR1 can bind dsRNA with any one of its three DRBDs, suggesting that these repeats are, in this regard, equivalent [167]. However, deletion of the first or third, but not the second DRBD, abolishes the editing activity. The importance of the first and third DRBD is supported by their conservation in ADAR2 and RED2, while the non-essential second DRBD lacks in ADAR2 and RED2 [156].

At the N-terminus, ADAR1, but not ADAR2 and RED2, contains the two domains Zα and Zβ, which specifically bind to left-handed Z-DNA [17]. The Zα domain is the first naturally occurring protein domain that binds Z-DNA with high affinity (K_d = 30 nM [4]), whereas Zβ binds Z-DNA with significantly lower affinity (Alan Herbert, unpublished results). The primary sequence alignment of Zα and Zβ domains shows that a large number of residues are conserved between human, mouse, rat, bovine and *Xenopus* ADAR1 (fig. 17). In particular, the three residues, L176, P192 and W195, (residues numbers of human ADAR1) are absolutely conserved throughout all sequences suggesting that they play an important role for the biological function of these protein domains. In the N-terminal third of the sequence alignment, the Zα sequences are more distinct from the Zβ and viral sequences suggesting that this part of the Zα domain may differ structurally and functionally from the Zβ and viral equivalent.

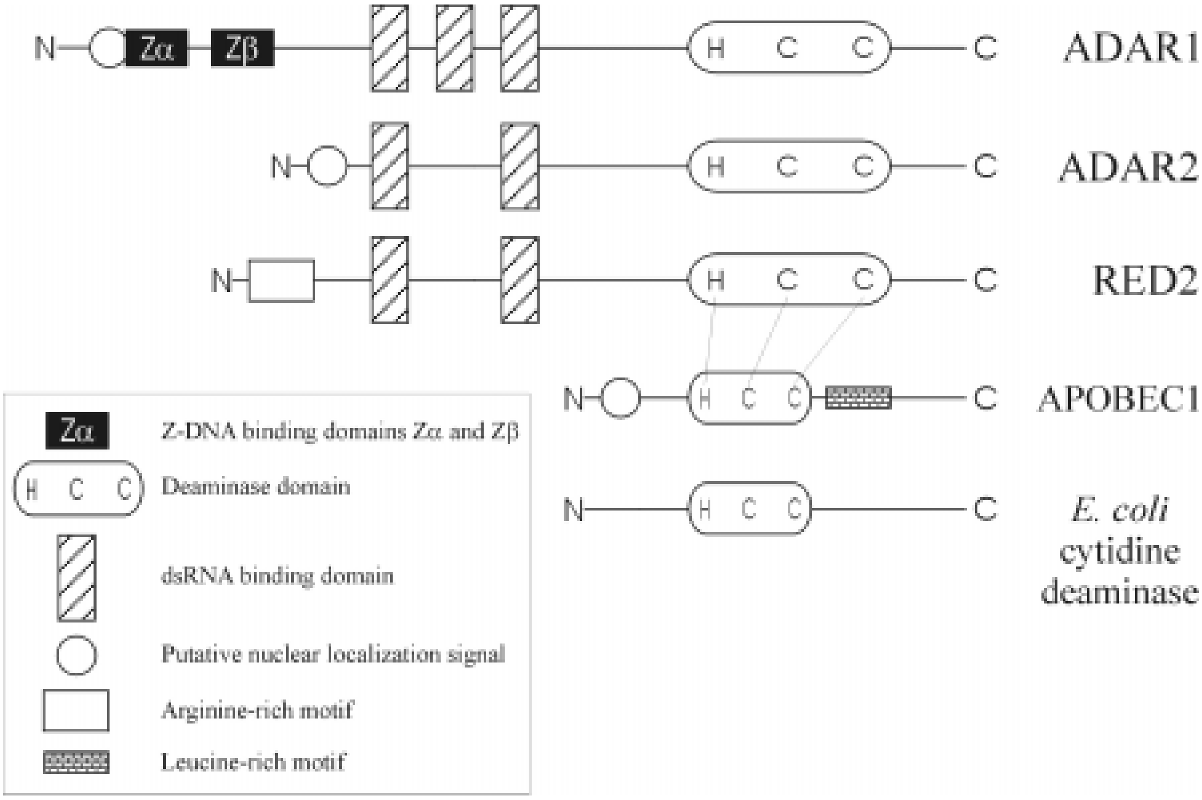

fig. 16 **Domain structure of mammalian RNA editing enzymes** (adapted from [1]). ADAR1, ADAR2, RED2 and APOBEC1 contain a homologous zinc-dependent deaminase domain that is similar to the cytidine deaminase of *E.coli*. The conserved residues, one histidine (H) and two cysteines (C), involved in Zn^{2+} coordination, are shown. The ADAR enzymes also contain two homologous dsRNA binding domains. But only ADAR1 possesses the N-terminal Z-DNA binding domains, Zα and Zβ.

Additionally, the sequences of the homologous domain of the E3L protein from vaccinia virus and its equivalent from variola virus were added. The E3L protein is of interest because it contains a dsRNA binding domain that is similar to those of the ADAR enzymes. E3L takes part in a viral defense mechanism counteracting the interferon-mediated anti-viral response of infected mammalian cells by binding to dsRNA. Thus, E3L resembles ADAR1 in its putative Z-DNA binding domain and the dsRNA binding domain, but E3L lacks a deaminase domain.

```
hZα   FQELSIYQD.QEQRILKFLEELGEGK.ATTAHDLSGKLGTPKK.EINRVLYSLAKKGKLQKEAGTPPLWKI
mZα   FRELSISQS.PEQKVLNRLEELGEGK.ATTAHVLARELRIPKR.DINRILYSLEKKGKLHRGRGKPPLWSL
rZα   FQELSISQN.PEQKVLNRLEELGEGK.ATTAYALARELRTPKK.DINRILYSLERKGKLHRGVGKPPLWSL
bZα   FQGLTISQD.QEQRTLELLDELGDGK.ATTARDLARKLQAPKK.DINRVLYSLAEKGKLHQEAGSPPLWRA
xZα1  FGSLTVSHDILENNLLTFFKEIG.TK.TFTAKALAWQFKVEKK.RINHFLYTFETKGLLCRYPGTPPLWRV
xZα2  FGSLSVSRDPLENILLTFFRGQGDTQ.TFTAKALAWQFKVKKK.HINYFLYKFGTKGLLCKNSGTPPLWKI
hZβ   LEFLDMAE-.IKEKICDYLFNVS--D.-SSALNLAKNIGLTKARDINAVLIDMERQGDVYRQGTTPPIWHL
mZβ   SEPLDMAE-.IKEKICDYLFNVS--N.-SSALNLAKNIGLTKARDVTSVLIDLERQGDVYRQGATPPIWYL
rZβ   SELLDMAE-.IKEKICDYLFNVS--K.-SSALNLAKNIGLAKARDVNAVLIDLERQGDVYREGATPPIWYL
bZβ   PEPLDMAE-.IKEKICDHLFNVSS--.-SSALNLAKNIGLTKARDVNAVLIDLERQGDVYRQGTTPPIWYL
xZβ1  CSPEDMAG-.NKEKVCEFLYNSPP--.-STTLIIRKNVGISKLPELNQILNTLEKQGEACKASTNPVKWTL
xZβ2  SSPEDMAT-.NSAKVCEFLYNSPP--.-STPFIIRKNVGISKMPELTQILNTLEKQGEACKASTNPVKWTL
e3l   MSKIYIDERSNAEIVCEAIKTIGIEG.-ATAAQLTRQLNMEKR EVNKALYDLQRSAMVYSSDDIPPRWFM
var   MSKIYIDERSDAEIVCEAIKNIGLEG.-VTAVQLTRQLNMEKR EVNKALYDLQRSAMVYSSDDIPPRWFM
```

fig. 17 **Primary sequence alignment of the Z-DNA binding domains, Zα and Zβ,** of human (h), mouse (m), rat (r), bovine (b) and *Xenopus* (x) ADAR1 (adapted from [3] and updated). In addition, the homologous sequences of the vaccinia virus E3L protein (e3l) and the variola virus equivalent (var) are shown. Three residues, L176, P192 and W195 (bold), are absolutely conserved throughout all Zα, Zβ and viral sequences. Moreover, many other residues are conserved as hydrophobic, polar or charged amino acids. Residues 130-197 of Zα and residues 291-355 of Zβ (numbering from human ADAR1, GenBank# U10439) are shown.

RNA editing and Z-DNA

RNA editing in higher eukaryotes is a young field of research with only a small number of substrates having been identified so far. Potentially, many pre-mRNAs can form dsRNA segments and thus become potential targets for ADAR1/2. During evolution the huge pool of intronic sequences may have accumulated the information necessary for double-strand formation with exons that labels these exonic sequences for ADAR-mediated editing. Thus, a larger number of yet unknown pre-mRNAs may be modified by RNA editing. The advent of large-scale genomic and cDNA sequencing projects promises the revelation of new RNA editing substrates.

In addition to the dsRNA binding domains and the deaminase domain, which are common to all ADAR enzymes, ADAR1 contains a unique Z-DNA binding domain, Zα, whose function is yet-to-be determined. The Zα domain may assist in the regulation of ADAR1-mediated editing as described by the following model [61]. When RNA polymerase transcribes a gene, negative supercoiling builds up at its 5' end causing short segments of alternating purine/pyrimidine sequences to transiently alter their conformation from B- to Z-DNA. Z-DNA segments of 6 bp or more are binding sites for the Zα domain [4], thereby targeting ADAR1 to the vicinity of the transcription complex. The co-localization with moving RNA polymerases ensures that the dsRNA binding domains of ADAR1 can bind to the emerging nascent pre-mRNA, and the deaminase domain can edit target sites before splicing occurs (fig. 18). After and during transcription, topoisomerase I relaxes the superhelical stress causing the Z-DNA segments to revert to B-DNA.

In this hypothetical model, the Zα/Z-DNA interaction confers editing site selectivity because transcripts of genes that contain Z-DNA forming sequences with a suitable spacing to the transcription complex are preferentially edited. An analysis of 137 fully sequenced human genes showed that such potentially Z-DNA forming sequences cluster in the 5' regions of these genes [84] suggesting that potential Zα binding sites are close to transcription initiation sites. Consequently, the Z-DNA binding activity of ADAR1 may help to overcome the problem of promiscuous ADAR1 editing observed in vitro.

Two experimental findings are consistent with this model. Firstly, the model mechanism ensures that editing take place before the pre-mRNA is processed or otherwise shielded. Secondly, introns (ECS) far upstream are capable of forming dsRNA editing sites, while introns downstream are yet-to-be synthesized. All distant ECS identified so far, reside in upstream regions.

A recent study showed that the two *Xenopus* ADAR1 variants, ADAR1.1 and ADAR1.2, localize to the nascent ribonucleoprotein matrix in transcriptionally active Lampbrush chromosomes in the nucleus of *Xenopus* oocytes [168]. ADAR1 was not distributed uniformly but particularly strong at a special chromosome loop. The addition of various inhibitors of splicing demonstrated that the localization of ADAR1 is independent of splicing which is consistent with the notion that ADAR-mediated editing occurs before splicing. Furthermore, the localization of ADAR1 was not affected by the deletion of the Z-DNA binding domain and the putative N-terminal nuclear localization domain indicating that the Zα domain is not generally required for targeting ADAR1 to transcriptionally active Lampbrush chromosomes. However, this finding does not exclude that the Zα domain is necessary for enzyme function *in vivo*.

Taken together, the Zα domain of ADAR1 plays a yet-to-be determined role in RNA editing *in vivo*. The hypothetical regulation model provides a plausible mechanisms of action for Zα in ADAR1-mediated RNA editing. Furthermore, it may lead to the discovery of a general mechanisms for linking biological processes, such as RNA processing, to transcription through protein/Z-DNA interactions. With this goal in mind, the structural and functional aspects the Zα/Z-DNA interaction have been investigated in this Ph.D. thesis and are presented in the following chapters.

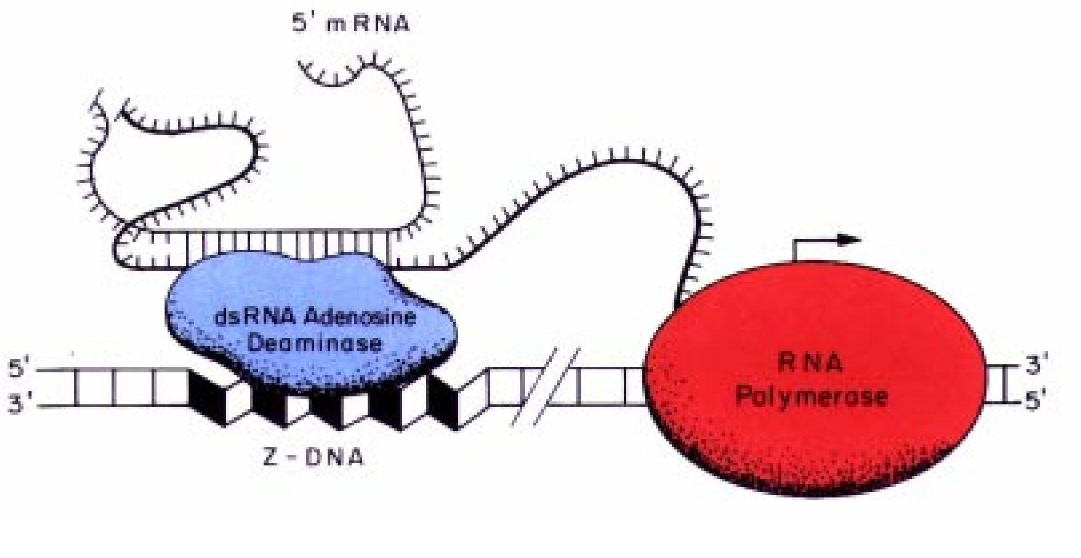

fig. 18 **A model for the regulation of ADAR1-mediated RNA editing by Z-DNA binding** (adapted from [61]). The moving RNA polymerase introduces negative supercoiling at its 5’ end (to its left side in this illustration) causing short segments with certain alternating purine/pyrimidine sequences to flip transiently into the Z-DNA conformation. The Z-DNA segments provide a binding site for Zα, thereby localizing the catalytic domain of ADAR1 next to the transcription complex where the substrate, nascent pre-mRNA, emerges.