

DNA Recognition by Eucaryotic Transcription Factors

DNA Recognition by Eucaryotic Transcription Factors

- The regulation of transcription in eucaryotes is in general much more complex and currently less well understood than the rather simple switch mechanisms that regulate procaryotic gene expression.
- There are three classes of RNA polymerases in eucaryotes.
- RNA polymerase I and RNA polymerase III transcribe the genes encoding ribosomal RNAs and transfer RNAs, respectively.
- RNA polymerase II transcribes genes that code for the messenger RNAs of proteins, and it is these genes that will be of our principal focus.

DNA Recognition by Eucaryotic Transcription Factors

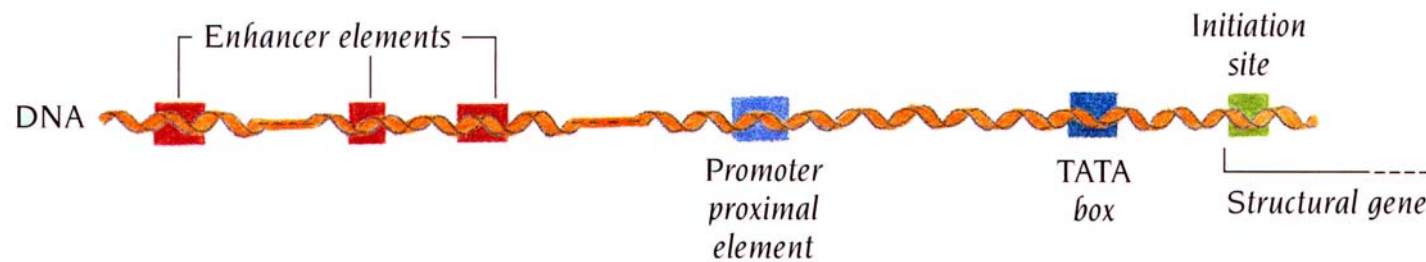
- Complex sets of regulatory elements control the initiation of transcription of these structural genes.

- Distal to the RNA polymerase II initiation site there are different combinations of specific DNA sequences, each of which is recognized by a corresponding site-specific DNA-binding protein. These proteins are called **transcription factors**, and each combination of DNA sequence and cognate transcription factor constitutes a **control module**.

- The essence of transcriptional regulation in eucaryotes is to use different combinations of a large set of control modules to regulate the expression of each gene.

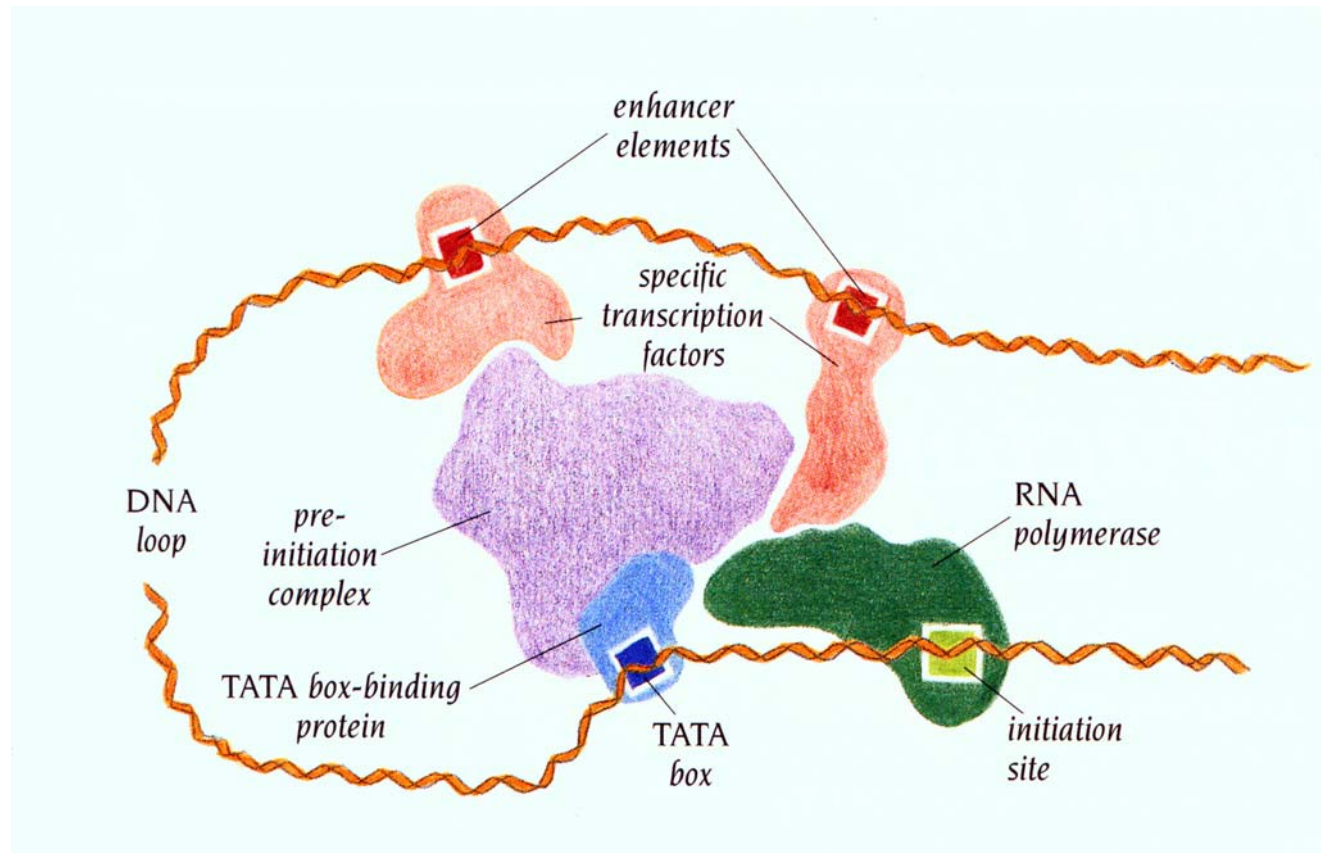
- Given the number of modules that is now being discovered in genomes of higher eucaryotes, the number of possible combinations is almost unlimited.

The transcriptional elements of a eucaryotic structural gene extend over a large region of DNA



The regulatory sequences can be divided into three main regions: (1) the basal promoter elements such as the TATA box, (2) the promoter proximal elements close to the initiation site, and (3) distal enhancer elements far from the initiation site. The promoter proximal elements are usually 100 to 200 base pairs long and relatively close to the site of initiation of transcription. Enhancer elements by contrast are short DNA sequences that occur further upstream or downstream from the initiator site than the proximal elements. They are often a few thousand base pairs from the promoter but may be 20,000 base pairs or more distant.

Schematic model for transcriptional activation



The TATA box-binding protein, which bends the DNA upon binding to the TATA box, binds to RNA polymerase and a number of associated proteins to form the pre-initiation complex. The complex interacts with different specific transcription factors that bind to promoter proximal elements and enhancer elements.

TATA box element and TATA box-binding protein

- The TATA box element is the best characterized core promoter element; a DNA sequence rich in A-T base pairs and located 25 base pairs upstream of the transcription start site.
- The TATA box is recognized by one of the basal transcription factors, the TATA box-binding protein, TBP, which is a part of a multisubunit complex called TFIID.
- TFIID in combination with RNA polymerase II and other basal transcription factors such as TFIIA and TFIIB form a preinitiation complex for transcription.

Transcription is activated by protein-protein interactions

- TFIID is believed to be the key link between specific transcription factors and the general preinitiation complex.
- TFIID contains the TATA box-binding protein in combination with a variety of different proteins called TBP-associated factors, TAFs.
- When the preinitiation complex has been assembled, strand separation of the DNA duplex occurs at the transcription start site, and RNA polymerase II is released from the promoter to initiate transcription. However, TFIID can remain bound to the core promoter and support rapid reinitiation of transcription by recruiting another molecule of RNA polymerase.

The TATA box-binding protein is ubiquitous

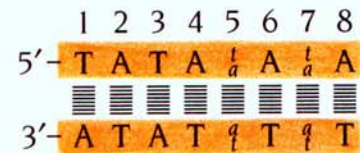
- TBP was first isolated and purified from yeast in 1988 and found to be composed of a single polypeptide chain with a molecular weight of 27 kDa.

- Comparison of TBPs of various organisms reveals that they are composed of a highly conserved C-terminal domain of 180 amino acids and an N-terminal domain that varies in length and shows little or no sequence conservation among different classes of organisms.

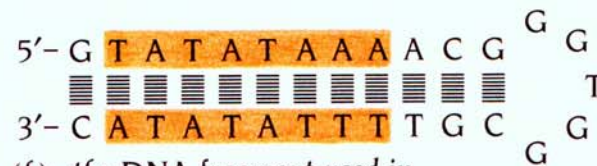
- TBP mutants lacking the N-terminal region are fully functional in promoter binding and stimulation of basal transcription and therefore these two functions must be provided by the C-terminal domain.

- The C-terminal domain of TBP contains all the functions essential for normal yeast growth and for responses to specific transcription activators with a net negative charge.

TATA box sequences



(a) consensus sequence of the TATA box



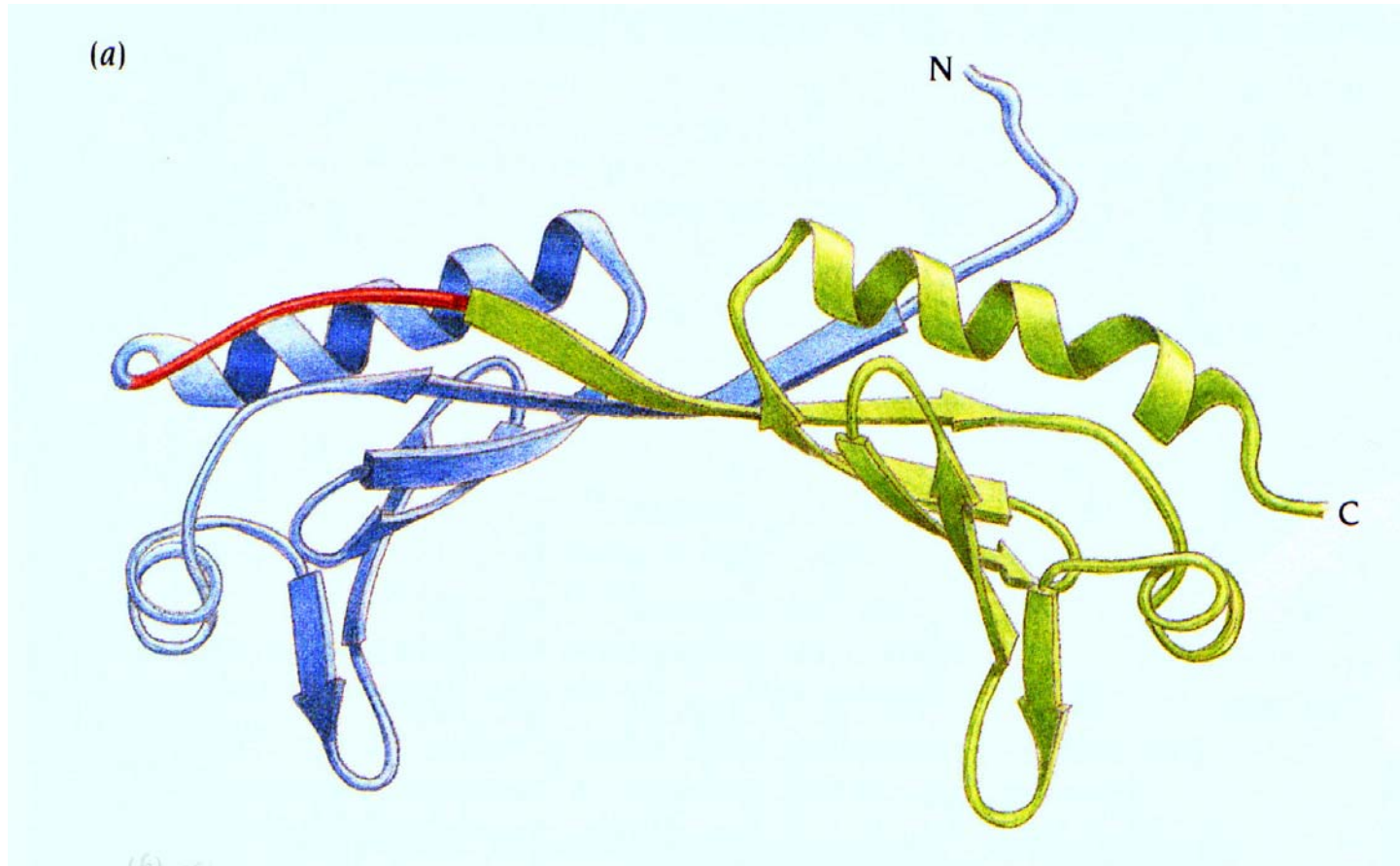
(b) the DNA fragment used in crystals of the complex with yeast TBP



(c) the DNA fragment used in crystals of the complex with *Arabidopsis thaliana* TBP

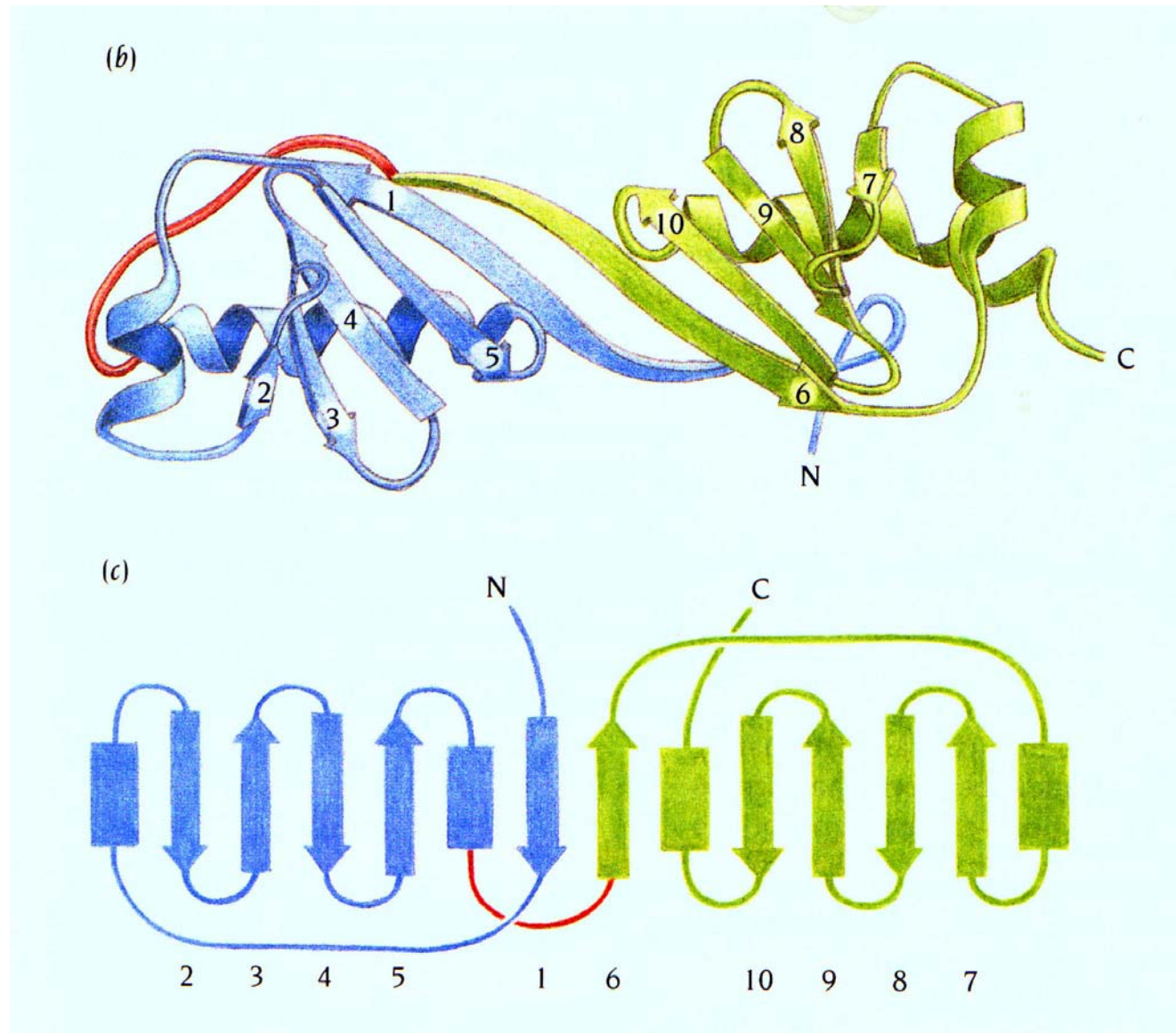
Comparison of the consensus nucleotide sequence of the TATA box (a) and the sequences of the DNA fragments used in the crystal structure determinations of the TATA box-binding proteins from yeast (b) and the plant *Arabidopsis thaliana* (c).

Structure of the TATA box-binding protein (TBP)

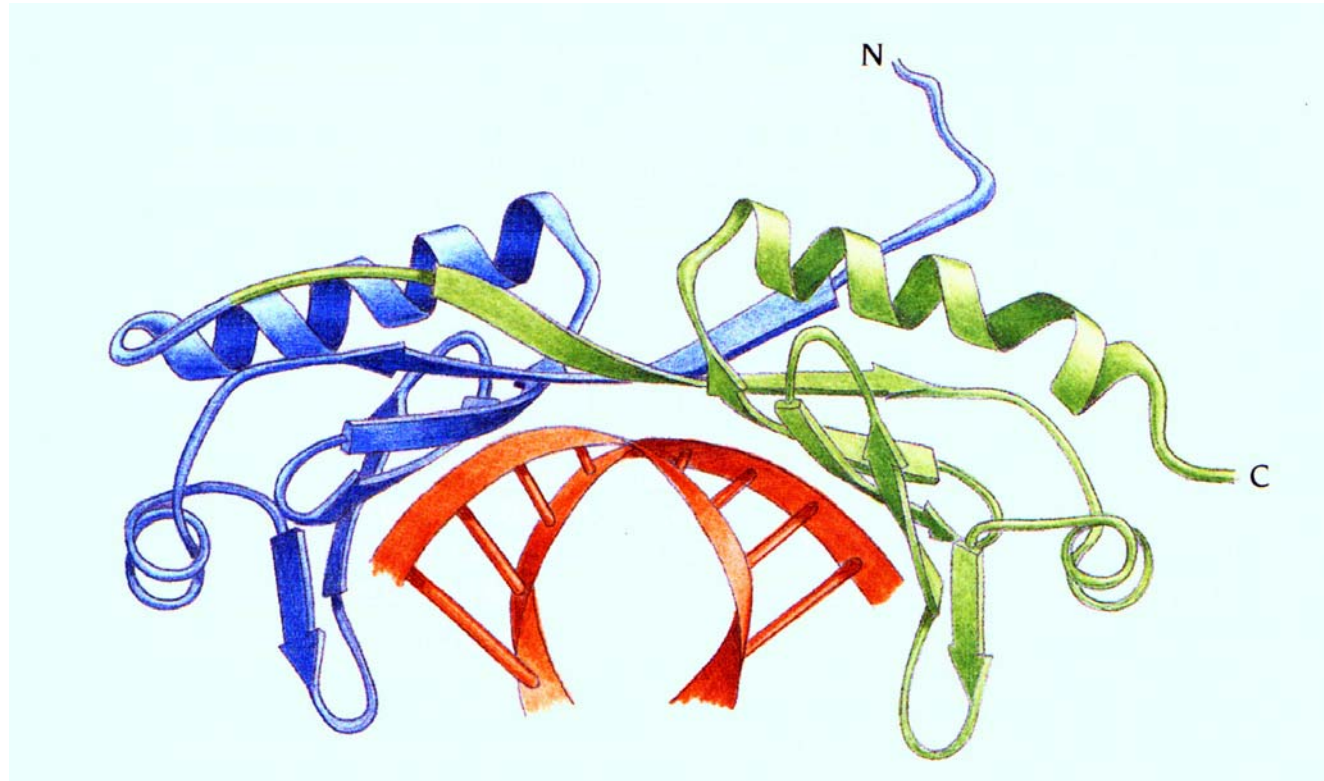


The two homologous repeats, each of 88 amino acids, at both ends of the TBP DNA-binding domain form two structurally very similar motifs. The two motifs each comprise an antiparallel β sheet of five strands and two helices. These motifs are joined together by a short loop to make a 10-stranded β sheet which forms a saddle-shaped molecule.

Structure of the TATA box-binding protein (TBP)



TBP binds in the minor groove and induces large structural changes in DNA



Side chains from the underside of the saddle and loops connecting β strands 2 and 3 form the DNA binding site. No α helices are involved in the interaction area, in contrast to the situation in most other eucaryotic transcription factors.

The DNA fragment is sharply bent in the TATA box region



The helical axis of the DNA at each end of the TATA box form an angle of about 100° to each other, instead of the expected 180° if the DNA was not bent.

The interaction area between TBP and the TATA box is mainly hydrophobic

- The interaction area between the TBP and the minor groove of DNA is formed by two large, complementary surfaces with no water molecules between them.
- A surprisingly large amount of this area is hydrophobic, in contrast to protein-DNA interactions that involve the major groove, which are mainly hydrophilic and sometimes mediated by water molecules.
- In TBP, the side chains from the eight central β strands interact with both the phosphate sugar backbone and the minor groove of the eight nucleotides of the TATA box.
- 15 side chains projecting from the β strands make hydrophobic contacts with the sugar and bases of DNA.
- The phosphate groups are hydrogen bonded to Arg and Lys side chains at the edges of the interaction area.

The interaction area between TBP and the TATA box is mainly hydrophobic

-Studies of site-directed mutations have shown that burying polar side chains in a hydrophobic environment without satisfying their hydrogen-bonding requirements is energetically unfavorable.

-Clearly, large number of energetically favorable hydrophobic interactions between TBP and TATA-box DNA must compensate for these unfavorable interactions in order to achieve binding constants in the nanomolar range.

The interaction area between TBP and the TATA box is mainly hydrophobic

-The only DNA sequence-specific aspect of these hydrophobic contacts is that they exclude G-C base pairs that are absent from the TATA box.

-The amino group of guanine would, if present, sterically disrupt or modify the interface.

-Surprisingly, 12 of the 16 hydrogen bond acceptors of the minor groove (N3 of adenine and O2 of thymine) are buried in this interaction area without forming hydrogen bonds either to protein side chains or to water molecules.

The interaction area between TBP and the TATA box is mainly hydrophobic

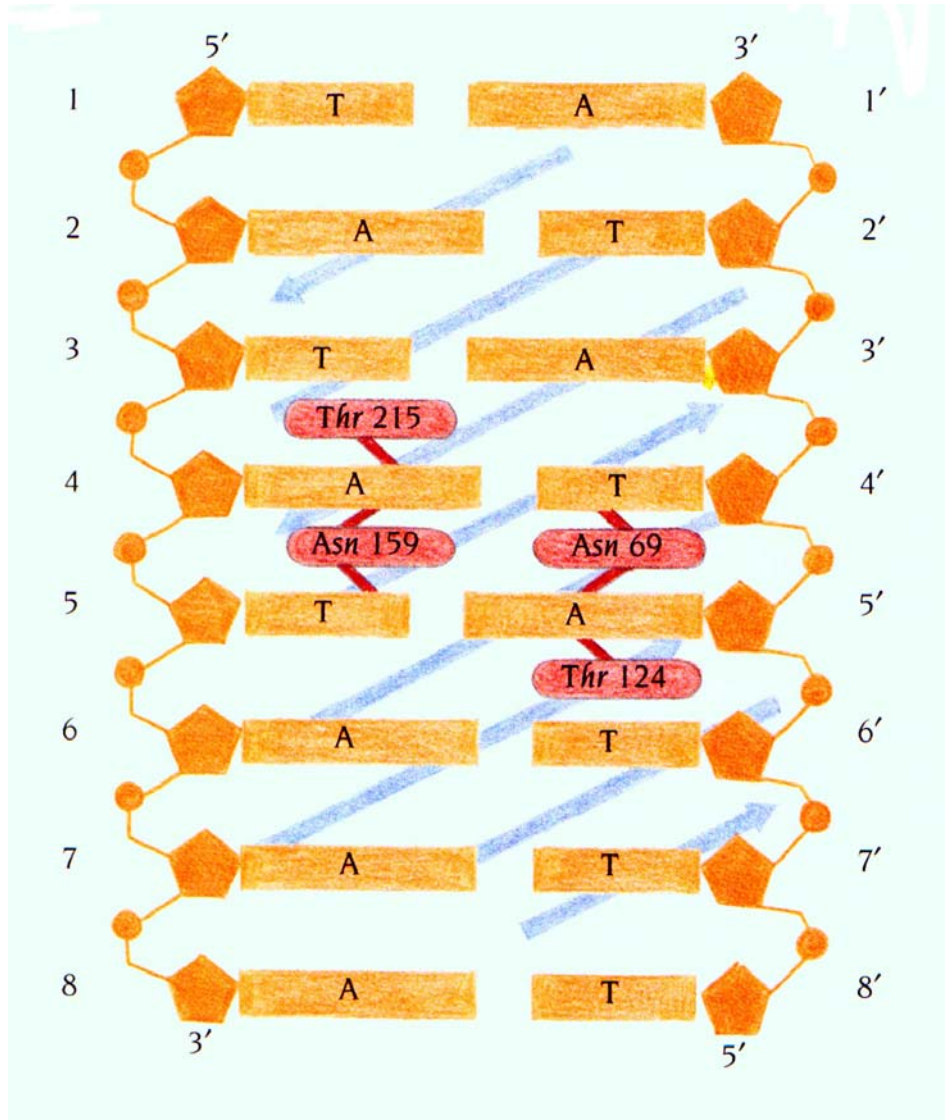
- The only sequence-specific hydrogen bonds between TBP side chains and the bases in the minor groove occur at the very center of the TATA box.

- The amide groups of two Asn side chains donate four hydrogen bonds, two each to adjacent bases on the same DNA strand (Asn 69 to O2 of T4' and to N3 of A5', and Asn 159 to O2 of T5 and N3 of A4).

- In addition, N3 of A5' and A4 accept hydrogen bonds from Thr 124 and Thr 215 respectively.

- There are also two conserved Val residues on each side of base pairs 4 and 5, Val 71 and 122 on one side and Val 161 and 213 on the other side. The side chains of the Val residues would cause steric interference with the NH₂ substituent from G-C or C-G base pair.

The interaction area between TBP and the TATA box is mainly hydrophobic



Sequence specific interactions between TBP and the TATA box by hydrogen bonds between the central region of the TATA box and Asn69 and Thr 124 from one domain and the equivalent residues Asn159 and Thr215 from the second domain.

The interaction area between TBP and the TATA box is mainly hydrophobic

-The flanking valine residues in combination with the six hydrogen bonds specify A-T or T-A at positions 4 and 5 of the TATA box.

The interaction area between TBP and the TATA box is mainly hydrophobic

-The consensus TATA-box sequence has an A-T base pair at position 4, but either a T-A or an A-T base pair at the symmetry-related position 5, and the sequence is, therefore, not strictly palindromic.

-However, the hydrogen bonds in the minor groove can be formed equally well to an A-T or a T-A base pair, because O2 of thymine and N3 of adenine occupy nearly stereochemically equivalent positions, and it is sufficient, therefore, for the consensus sequence of the TATA box to be quasi-palindromic.

The interaction area between TBP and the TATA box is mainly hydrophobic

-Like Thr 124 and Thr 215, the Asn 69 and Asn 159 residues occupy equivalent positions in the two homologous motifs of TBP.

-By analogy with the symmetric binding of a dimer repressor molecule to a palindromic sequence, the two motifs of TBP form symmetric sequence-specific hydrogen bonds to the quasi-palindromic DNA sequence at the center of the TATA box.

The interaction area between TBP and the TATA box is mainly hydrophobic

In conclusion:

- One important factor that contributes to the strong affinity of TBP proteins to TATA boxes is the large hydrophobic interaction area between them.

- Major distortions of the B-DNA structure cause the DNA to present a wide and shallow minor groove surface that is sterically complementary to the underside of the saddle structure of the TBP protein.

- The complementarity of these surfaces, and in addition the six specific hydrogen bonds between four side chains from TBP and four hydrogen bond acceptors from bases in the minor groove, are the main factors responsible for causing TBP to bind to TATA boxes 100,000-fold more readily than to a random DNA sequence.

Homeodomain proteins are involved in the development of many eucaryotic organisms

- Eucaryotes have many more genes and broader range of specific transcription factors than procaryotes and gene expression is regulated by using sets of these factors in combinatorial way.
- Eucaryotes have found several different solutions to the problem of producing a three-dimensional scaffold that allows a protein to interact specifically with DNA. Some of the solutions that have no counterpart in procaryotes will be discussed later.
- However, the procaryotic **helix-turn-helix** solution to this problem is also exploited in eucaryotes, in **homeodomain proteins** and some other families of transcription factors.

Homeodomain proteins are involved in the development of many eucaryotic organisms

-The homeobox, a DNA sequence of about 180 base pairs within the coding region of certain genes, was first discovered in the fruitfly *Drosophila* during studies of mutations that cause bizarre disturbances of the fly's body plan, so-called homeotic transformations.

-In the mutation *Antennapedia*, for example, legs grow from the head in place of antennae. Such homeotic mutations cause a whole set of cells to be misinformed as to their location in the organism and consequently to form a structure appropriate to another place.

Homeodomain proteins are involved in the development of many eucaryotic organisms

-Homeoboxes have since been found in several hundred different genes from both vertebrates and invertebrates, and there are varying degrees of DNA sequence homology between different members of this superfamily.

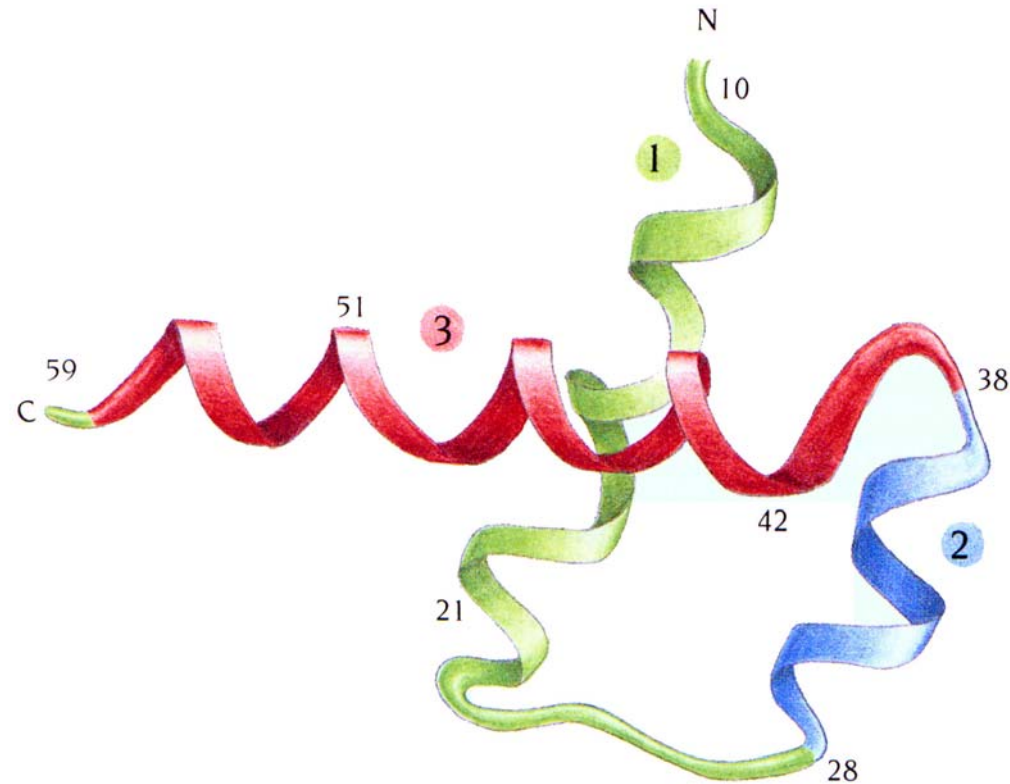
-A number of these homeobox genes retained both their precisely ordered tandem arrangement in the genome, and their developmental roles in axial patterning across more than 500 million years of evolution.

Homeodomain proteins are involved in the development of many eucaryotic organisms

-Homeoboxes code for homeodomains, sequences of 60 amino acids that function as the DNA-binding regions of transcription factors.

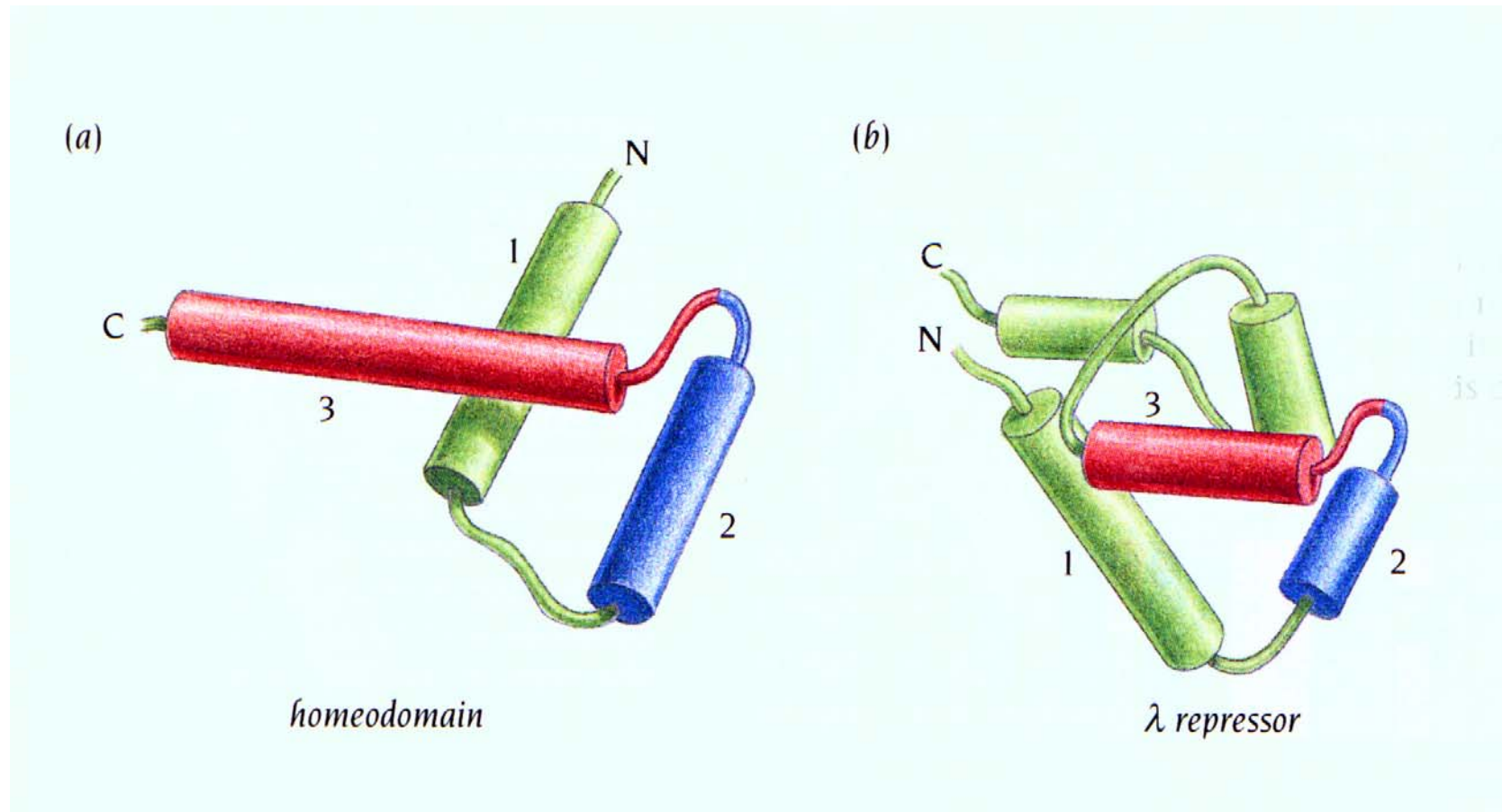
-Each homeobox gene of *Drosophila* is expressed only in its own characteristic subset of embryonic cells, and almost every embryonic cell contains a unique combination of homeodomain proteins.

Monomers of homeodomain proteins bind to DNA through a helix-turn-helix motif



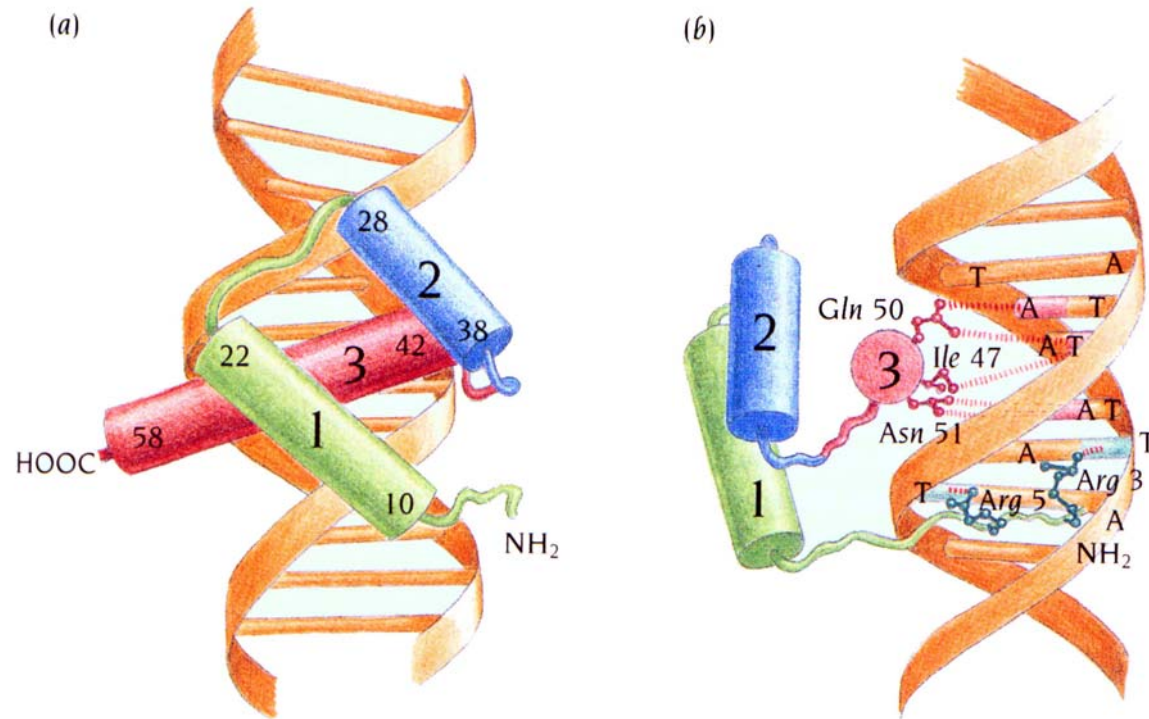
The structure of Antennapedia homeodomain is built up from three α helices connected by short loops. Helices 2 and 3 form a helix-turn-helix motif similar to those in procaryotic DNA-binding proteins.

Comparison of the helix-turn-helix motifs in homeodomains (a) and λ repressor (b)



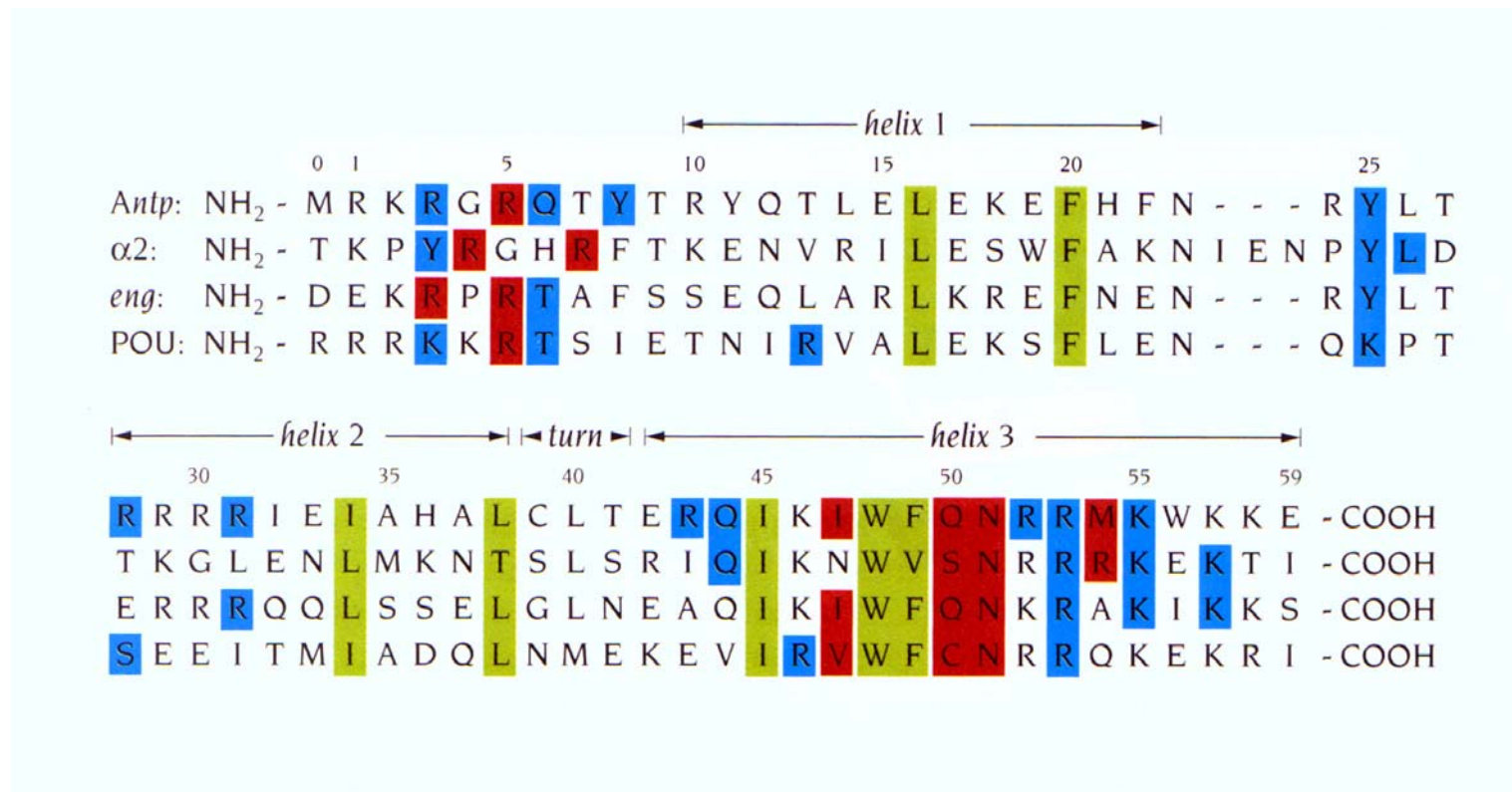
The recognition helix (red) of the homeodomain is longer than in the procaryotic repressor motif. In addition, the first helix (green in a) is oriented differently.

Complex between DNA and one monomer of the homeodomain



The recognition helix (red) binds in the major groove of DNA and provides the sequence-specific interactions with bases in the DNA. The N-terminus (green) binds in the minor groove on the opposite side of the DNA molecule and arginine side chains make nonspecific interactions with the phosphate groups of the DNA.

Amino acid sequences of homeodomains from four different transcription factors



Antp is from the *Antennapedia* gene in the fruitfly *Drosophila*, $\alpha 2$ is from the yeast *Mat $\alpha 2$* gene, eng is from the engrailed gene of *Drosophila* and POU is from the POU homeodomain in the mammalian gene *Oct-1*. Residues colored green form the hydrophobic core of the homeodomain, blue form nonspecific interactions with the DNA backbone and red form contacts with the edges of the DNA bases.

The positions of the residues involved in the nonspecific DNA contacts vary slightly in the homeodomain complexes but the overall arrangement of the homeodomains bound to DNA is virtually identical.

The amino acid sequence identity between the three is only 20%, and the remarkable conformational identity between such distantly related members of the homeodomain family strongly suggests that their conserved structure-function relationships have played important roles in the evolution of eucaryotic organisms.

***In vivo* specificity of homeodomain transcription factors depends on interactions with other proteins**

-Homeodomains that have closely related amino acid sequences bind with comparable affinities to the same DNA sequence *in vitro*. But *in vivo* transcription factors with similar homeodomains must activate the expression of their different target genes in a highly selective and precise way to avoid, for example, catastrophic deformations of the developing body plan of an organism.

-How do functionally different homeodomain proteins achieve necessary selectivity and specificity when their DNA binding regions are virtually identical?

-Homeodomain proteins, like other classes of transcription factors, do not operate alone but in concert with other transcription factors and associated proteins. These additional proteins cooperate with the homeodomain to provide the required specificity for the target genes.

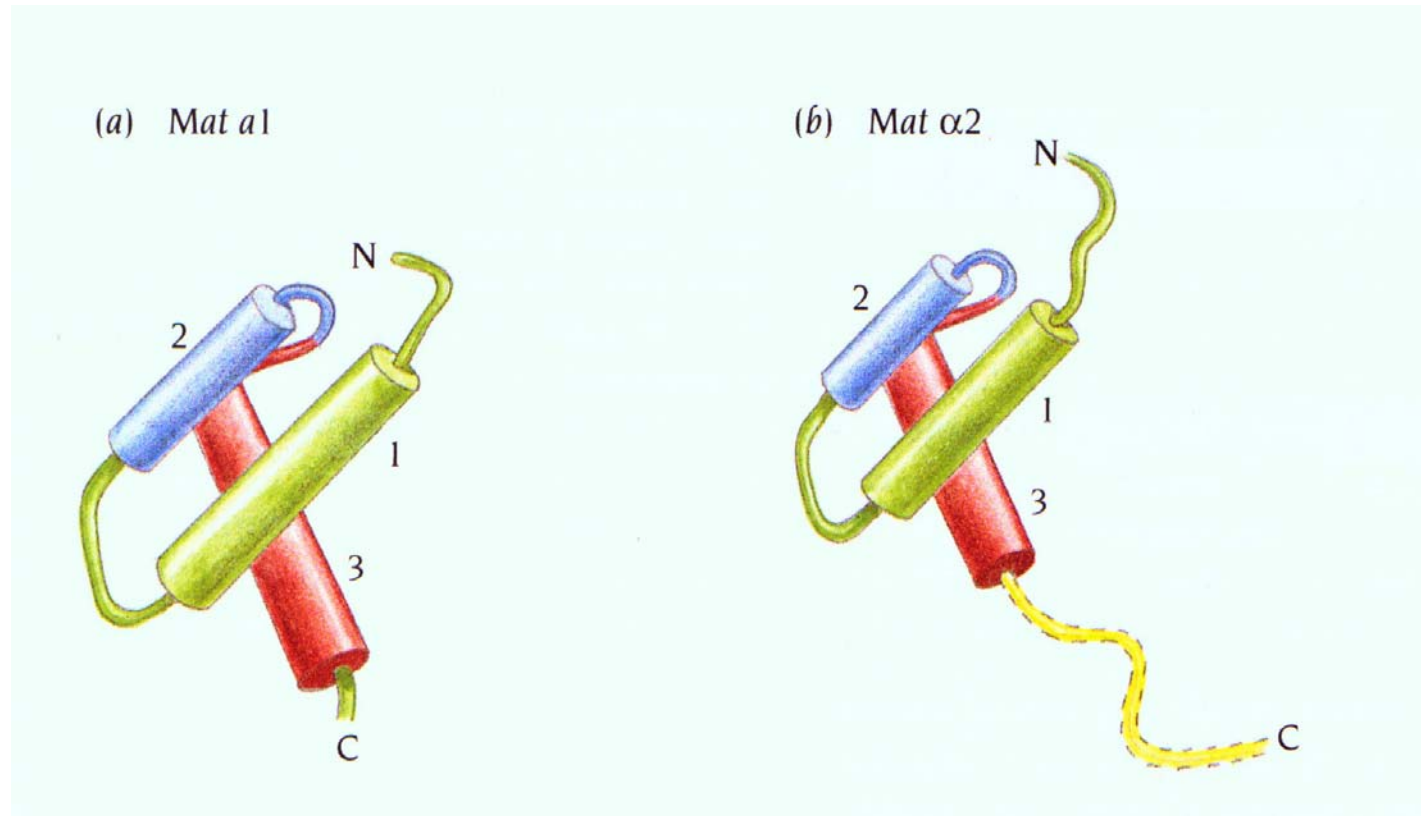
***In vivo* specificity of homeodomain transcription factors depends on interactions with other proteins**

-Combinations of the **Mat α 2** homeodomain transcription factor with either of two different transcription factors **Mcm1** or **Mat a1** specify two different cell types in yeast. Mat a1 has a homeodomain whereas Mcm 1 is a different type of transcription factor.

-Mat α 2 contains an N-terminal dimerization domain that binds to Mcm 1 and a C-terminal region that binds to Mat a1 in each case to form a DNA-binding heterodimer.

-How is the binding specificity of the heterodimer achieved compared with the specificity of Mat α 2 alone?

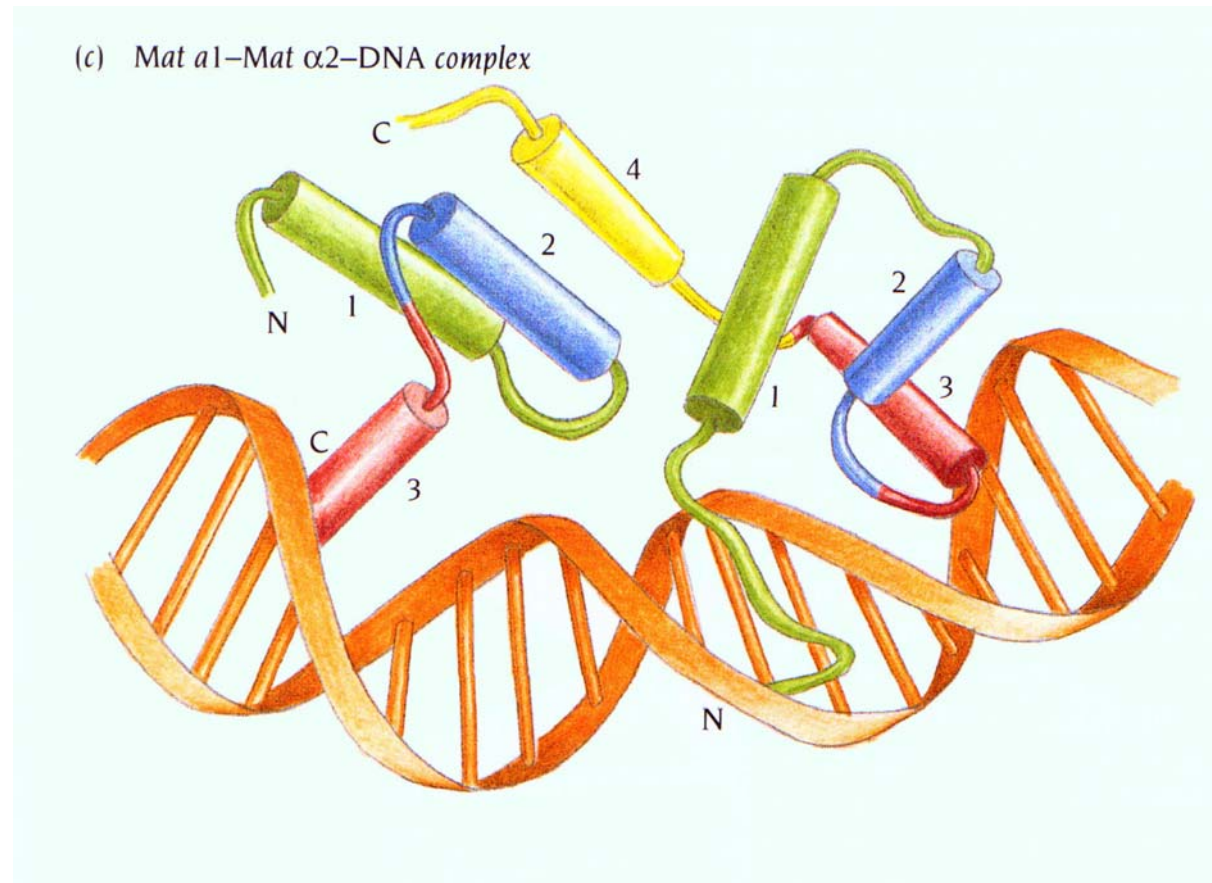
***In vivo* specificity of homeodomain transcription factors depends on interactions with other proteins**



(a) The assumed structure of the *Mat a1* homeodomain in the absence of DNA, based on its sequence similarity to other homeodomains of known structure.

(b) The structure of the *Mat α 2* homeodomain. The C-terminal tail (dotted) is flexible in the monomer and has no defined structure.

In vivo specificity of homeodomain transcription factors depends on interactions with other proteins



The structure of the heterodimeric yeast transcription factor Mat $\alpha 2$ – Mat $\alpha 1$ bound to DNA. The C-terminal domain of Mat $\alpha 2$ (yellow) folds into an α helix (4) in the complex and interacts with the first two helices of Mat $\alpha 1$, to form a heterodimer that binds to DNA.

***In vivo* specificity of homeodomain transcription factors depends on interactions with other proteins**

-The contacts between the Mat $\alpha 2$ homeodomain and DNA in the heterodimer complex are virtually indistinguishable from those seen in the structure of the Mat $\alpha 2$ monomer bound to DNA.

-However, there are at least two significant factors that may account for the increased specificity of the heterodimer.

***In vivo* specificity of homeodomain transcription factors depends on interactions with other proteins**

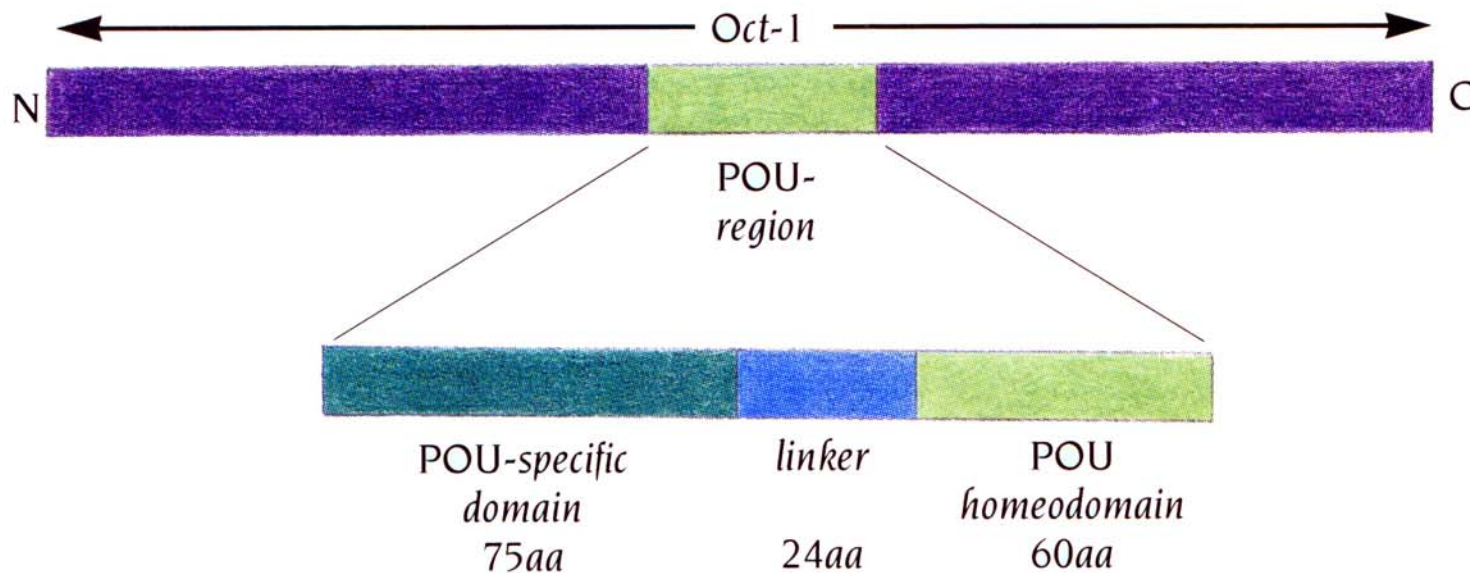
-First, the Mat a1 homeodomain makes significant contacts with the DNA, and the heterodimeric complex will therefore bind more tightly to sites that provide the contacts required by both partners.

-Second, site-directed mutagenesis experiments have shown that the protein-protein interactions involving the C-terminal tail of Mat $\alpha 2$ and the Mat a1 homeodomain are crucial for the specificity of binding. The heterodimer interface may increase specificity by dictating the precise spacing of the two binding sites for the recognition helices, analogous to the 434 Cro and repressor proteins discussed earlier. This could be one way in which regions of transcription factors outside the DNA-binding domains, the so-called activating regions, provide specificity of transcriptional activation *in vivo*.

POU DNA-binding region

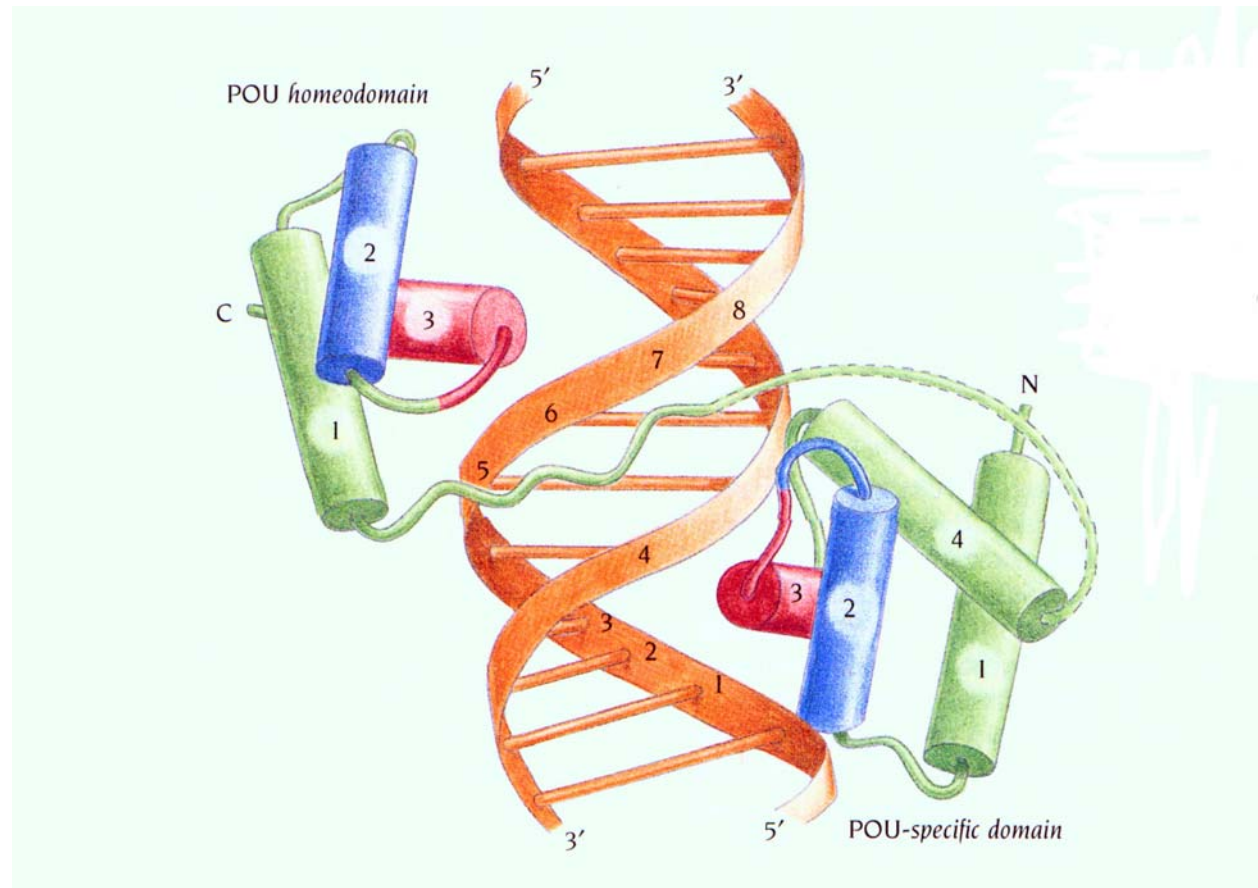
- A tandem arrangement of two DNA-binding domains, called the **POU region**, is present in a class of transcription factors involved in the expression of growth factors, histones and immunoglobulins.
- The POU region is a 150-160 amino acid sequence consisting of a homeodomain and a POU-specific domain joined by a short variable linker region.
- The POU-specific domain has a structure very similar to the λ repressor, in spite of having no apparent sequence homology; this structure is very simple consisting of four α helices where **helices 2 and 3 form a helix-turn-helix motif**.
- The POU region has, therefore, two **helix-turn-helix motifs** that are similar while the remaining structures of its two domains are quite different.

POU regions bind to DNA by two tandemly oriented helix-turn-helix motifs



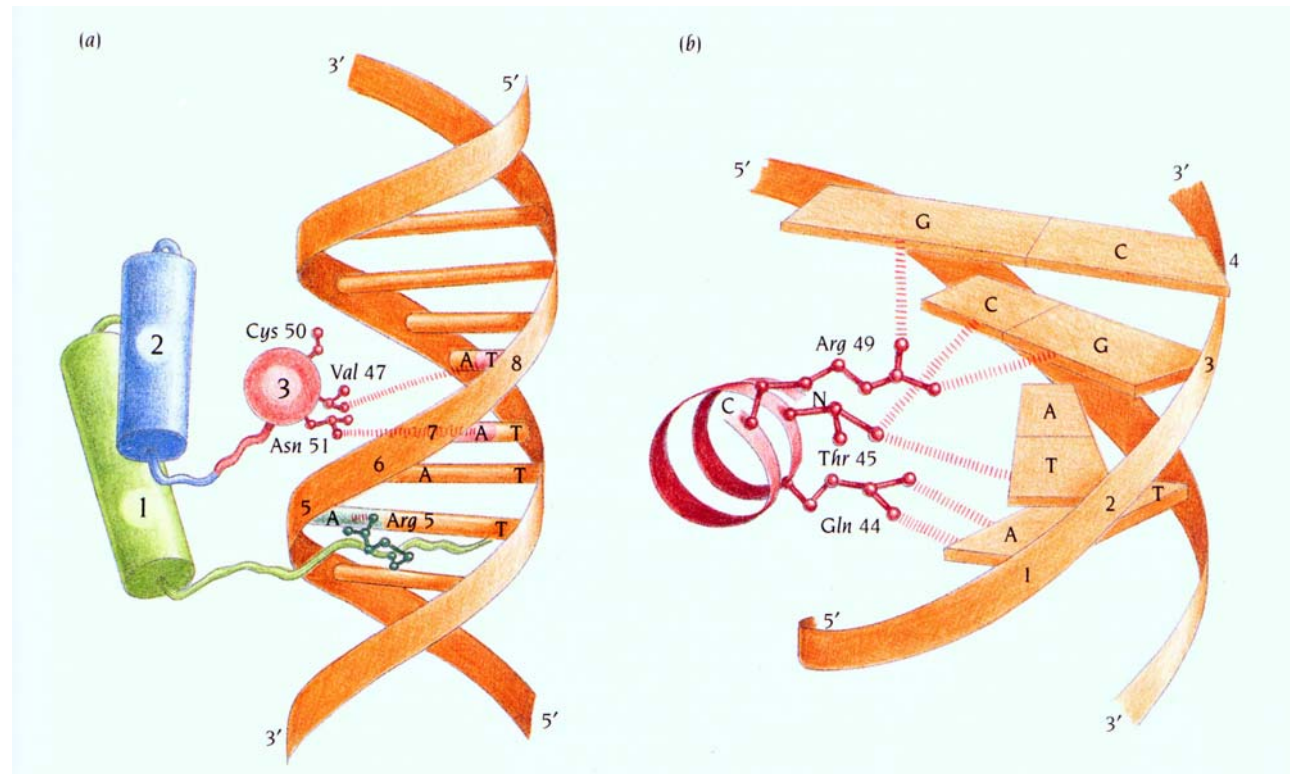
The DNA binding region of the protein Oct-1, the POU region, comprises two domains, the POU-specific domain and the POU homeodomain joined by a linker region. These two domains bind to DNA in a tandem arrangement.

POU regions bind to DNA by two tandemly oriented helix-turn-helix motifs



The two domains of the POU region bind in tandem on opposite sides of the DNA double helix. Both the POU-specific domain and the POU homeodomain have a helix-turn-helix motif (blue and red) which binds to DNA with their recognition helices (red) in the major groove. The linker region that joins these domains is partly disordered.

POU regions bind to DNA by two tandemly oriented helix-turn-helix motifs



The sequence specific contacts between DNA and the POU region. (a) Contacts from the **POU homeodomain**. Three residues from the recognition helix, **Val 47**, **Cys 50** and **Asn 51**, interact with base pairs 7 to 8 in the major groove of the second half of the POU-binding site. In addition, **Arg 5** in the N-terminal arm interacts with base pair 5 of the DNA in the minor groove. (b) Contacts from the **POU-specific region**. Three residues from the recognition helix, **Gln 44**, **Thr 45** and **Arg 49**, interact with bases in the major groove of the first half of the POU-binding site, base pairs 1 to 4. These contacts are from the opposite side of the DNA molecule compared with those in (a).

Understanding tumorigenic mutations

- The p53 protein plays a fundamental role in human cell growth and mutations in this protein are frequently associated with the formation of tumors.
- It is estimated that of the 6.5 million people diagnosed with one or another form of cancer each year about half have p53 mutation in their tumor cells and that the vast majority of these mutations are single point mutations.
- Wild-type p53** inhibits tumor formation, and its gene is therefore called a **tumor suppressor gene**. Details how this suppression is accomplished are currently unknown, but one important role of **p53** is to maintain the integrity of the genome during cell division by controlling a critical step in the cell cycle.

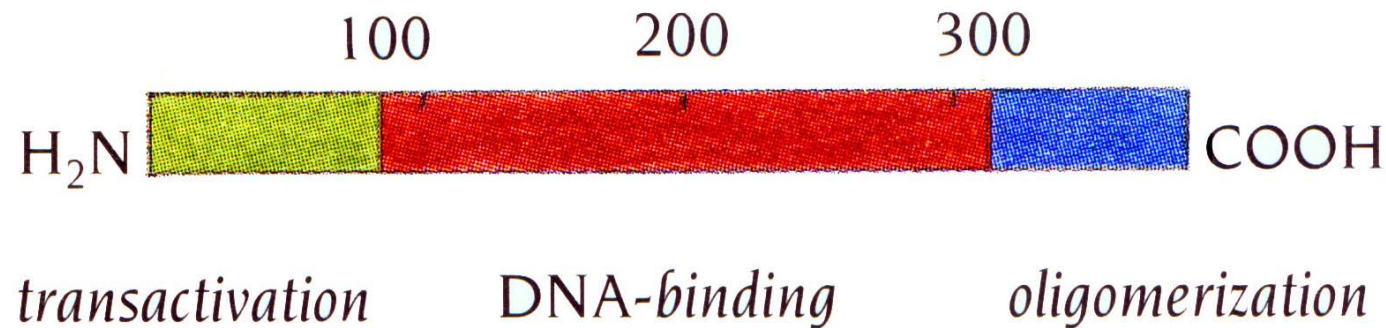
Understanding tumorigenic mutations

- A complex series of molecular interactions regulates the cell cycle; the chief players are proteins called cyclins and their associated enzymes, cyclin-dependent kinases.
- These kinases are inhibited by another protein, p21, whose expression is promoted by p53.
- In the presence of p21 the cell cycle is halted before the cell is committed to divide.
- Presumably this gives the cell time either to repair damaged DNA or, if it is beyond repair, to initiate programmed cell death (apoptosis).

Understanding tumorigenic mutations

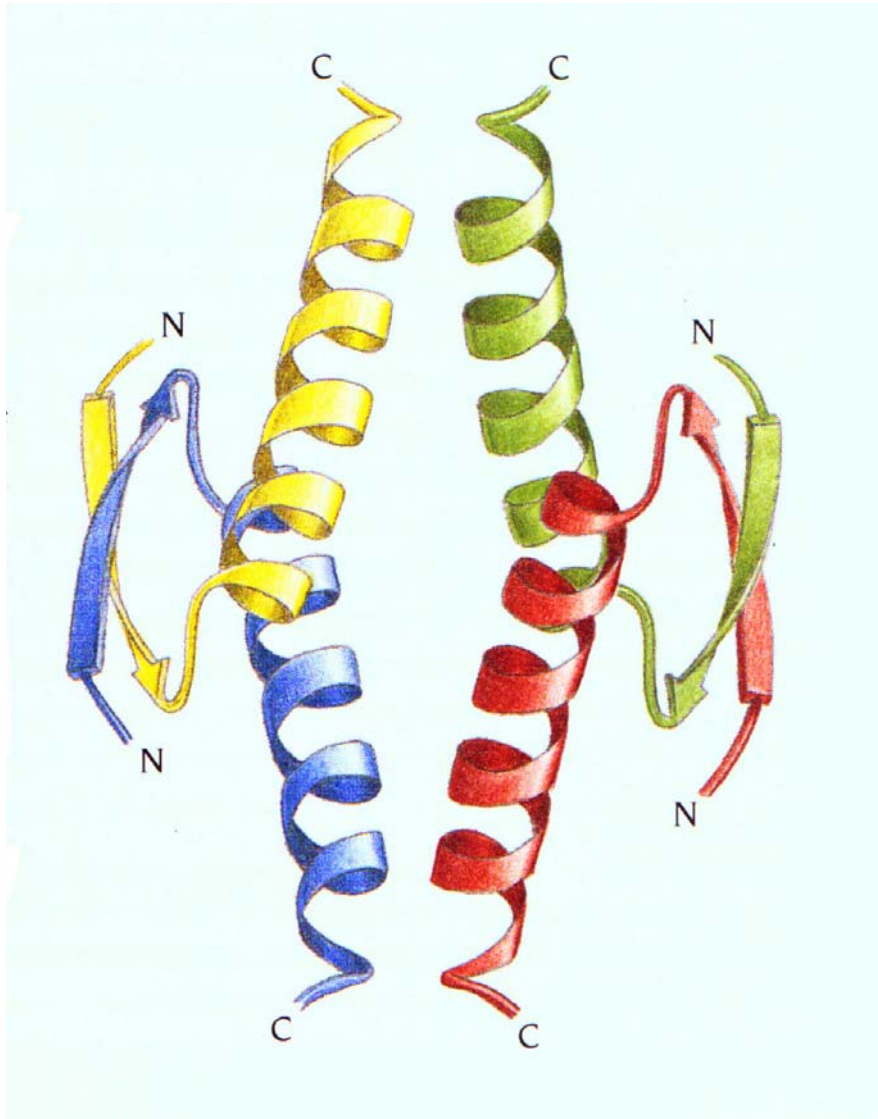
- One of the most important molecular functions of p53 is therefore to act as an activator of p21 transcription.
- The wild-type protein binds to specific DNA sequences, whereas tumor-derived p53 mutants are defective in sequence-specific DNA binding and consequently cannot activate the transcription of p53-controlled genes.
- More than half of the over one thousand different mutations found in p53 involve amino acids which are directly or indirectly associated with DNA binding.

The monomeric p53 polypeptide chain is divided in three domains



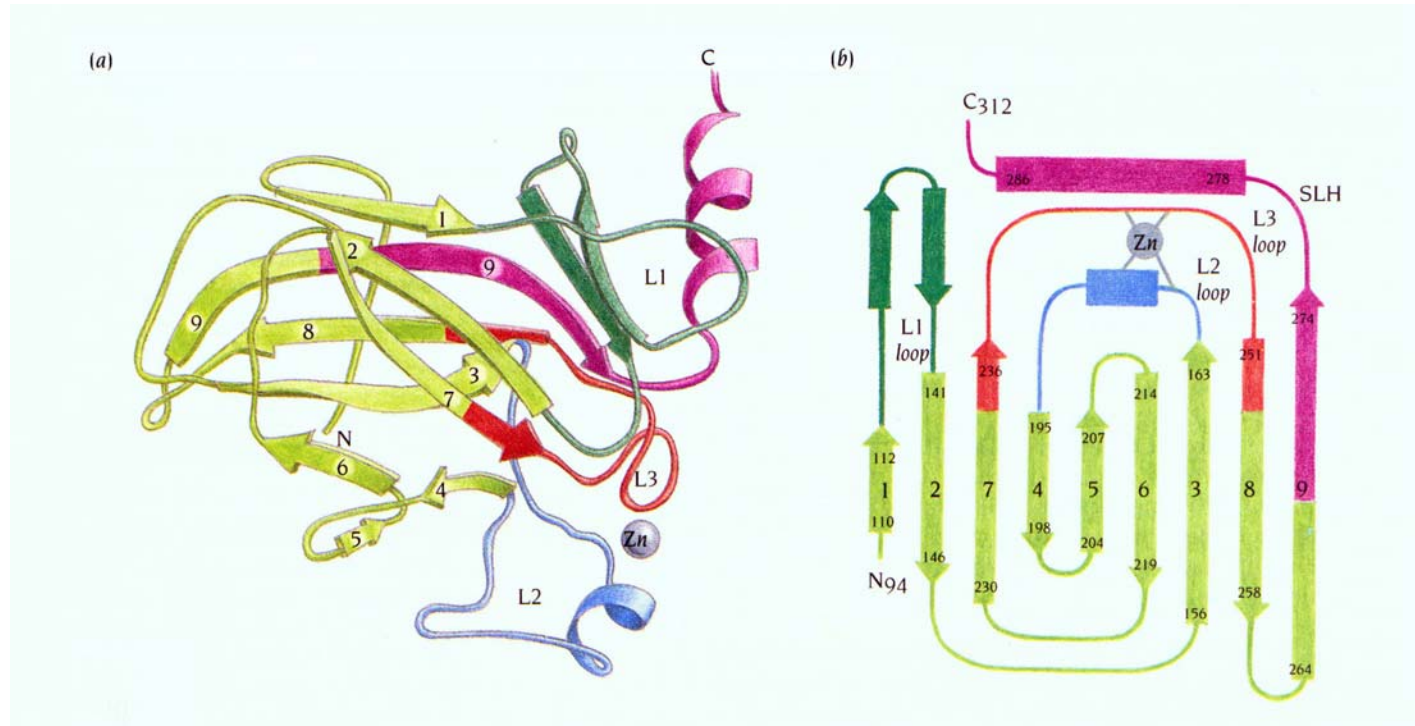
The polypeptide chain of the protein p53 is divided into three domains with different functions: transactivation, DNA binding and oligomerization.

The oligomerization domain of p53 forms tetramers



The four p53 subunits have different colors. Each subunit has a simple structure comprising a β strand and an α helix joined by a one-residue turn. The tetramer is built up from a pair of dimers (yellow-blue and red-green). Within each dimer the β strands form a two-stranded antiparallel β sheet which provides most of the subunit interactions. The two dimers are held together by interactions between the four α helices, which are packed in a different way from a four-helix bundle.

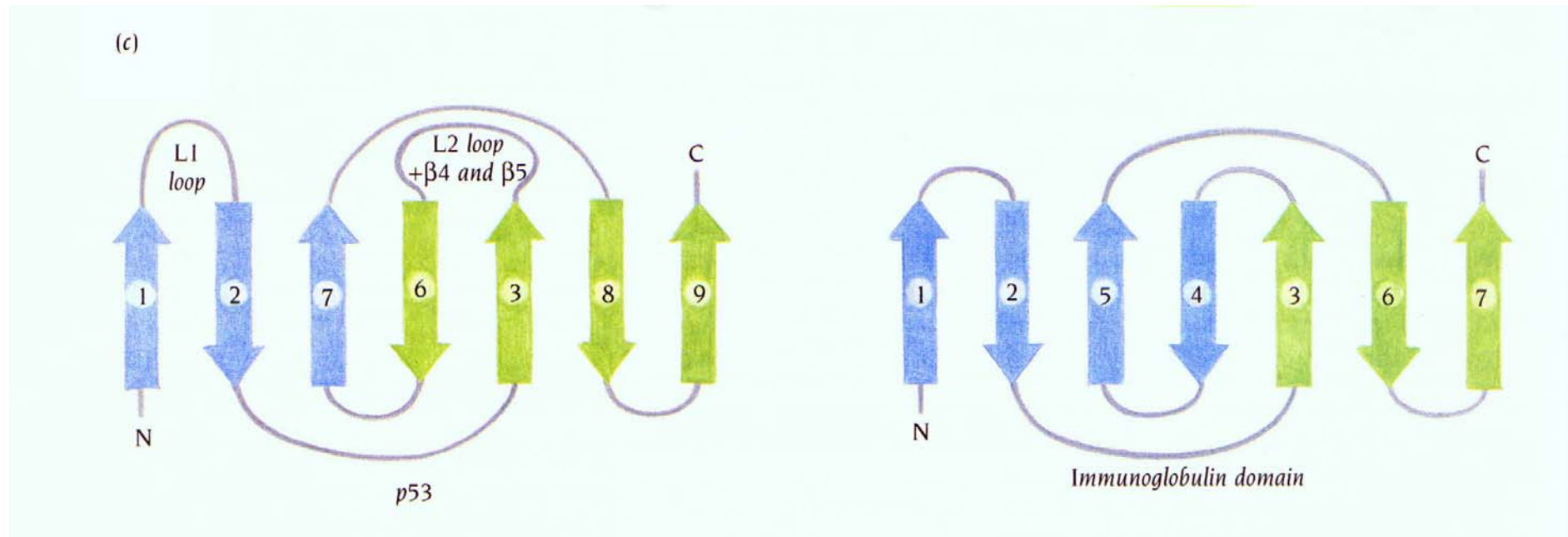
The DNA-binding domain of p53 is an antiparallel β barrel



(a) The DNA binding domain of p53 folds into an antiparallel β barrel with long loop regions – L1, L2 and L3 – at one end of the molecule. The conformations of loops L2 and L3 are stabilized by a zinc atom. The C-terminus of the chain (purple) is at the same end of the molecule as these loops.

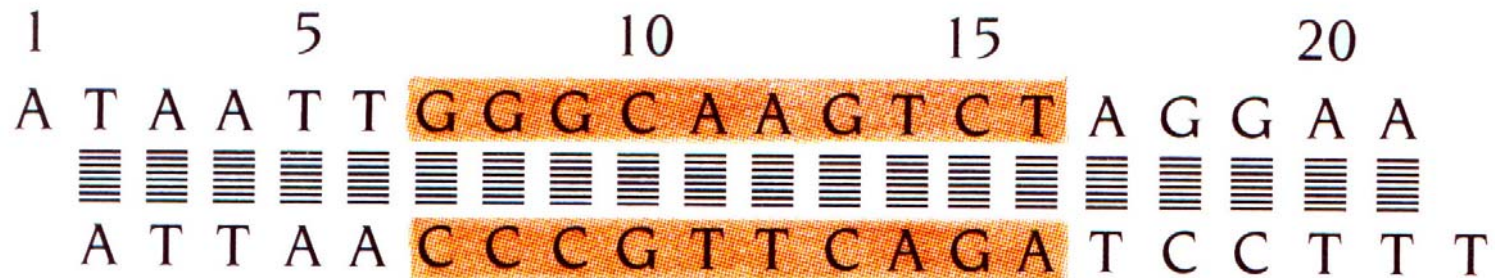
(b) Topological diagram of the DNA-binding domain of p53. Nine antiparallel β strands form a β barrel, the central part of which has a fold similar to that of immunoglobulin domains. A region called SLH (strand-loop-helix) spans residues 271-286.

The DNA-binding domain of p53 is an antiparallel β barrel



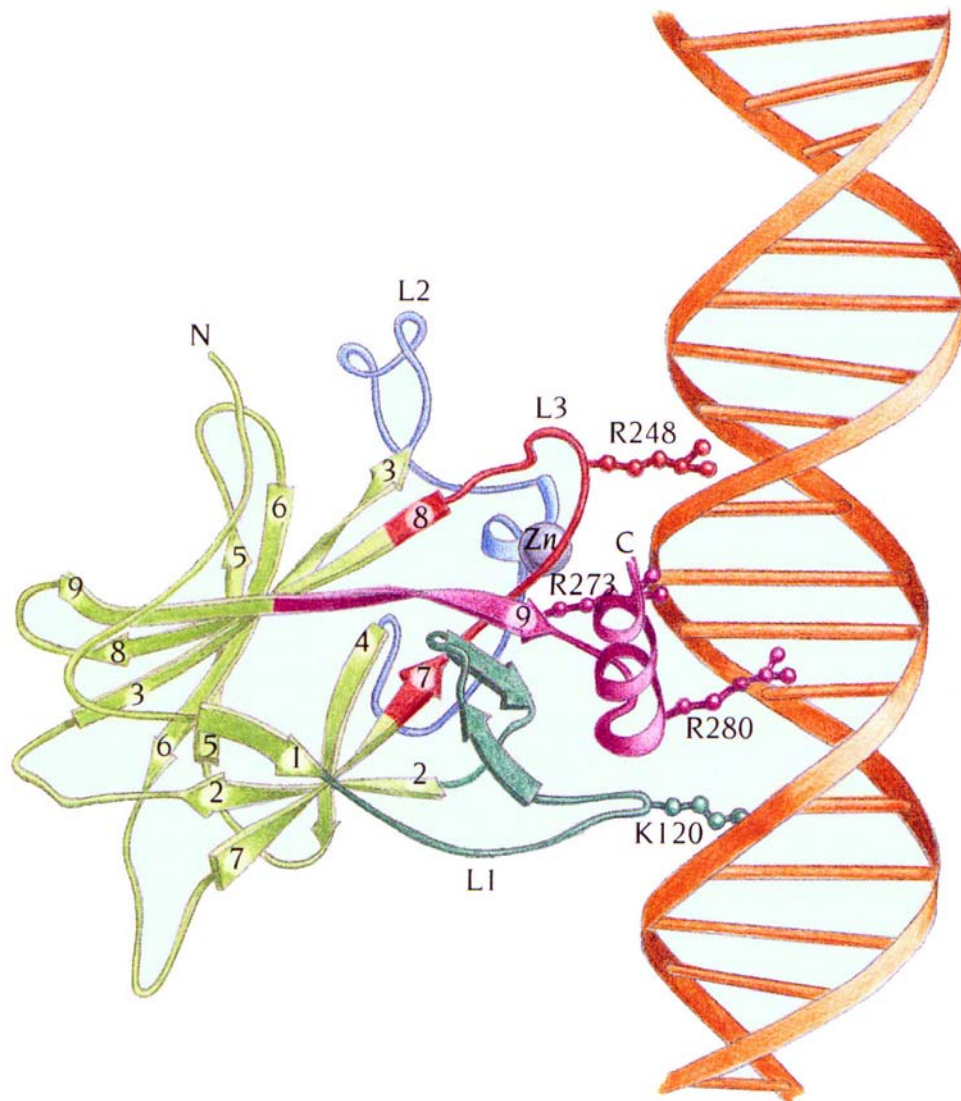
Simplified topological diagram of p53 compared with a topological diagram of the constant immunoglobulin domain. Strands of the same color belong to the same β sheet. Strands number 4 and 5 are considered to be part of the loop between strands 3 and 6. The middle strand in these diagrams belongs to different sheets but in the actual structures these strands are at the edges of the two sheets, in only slightly different positions relative to the common structure.

Two loop regions and one α helix of p53 bind to DNA



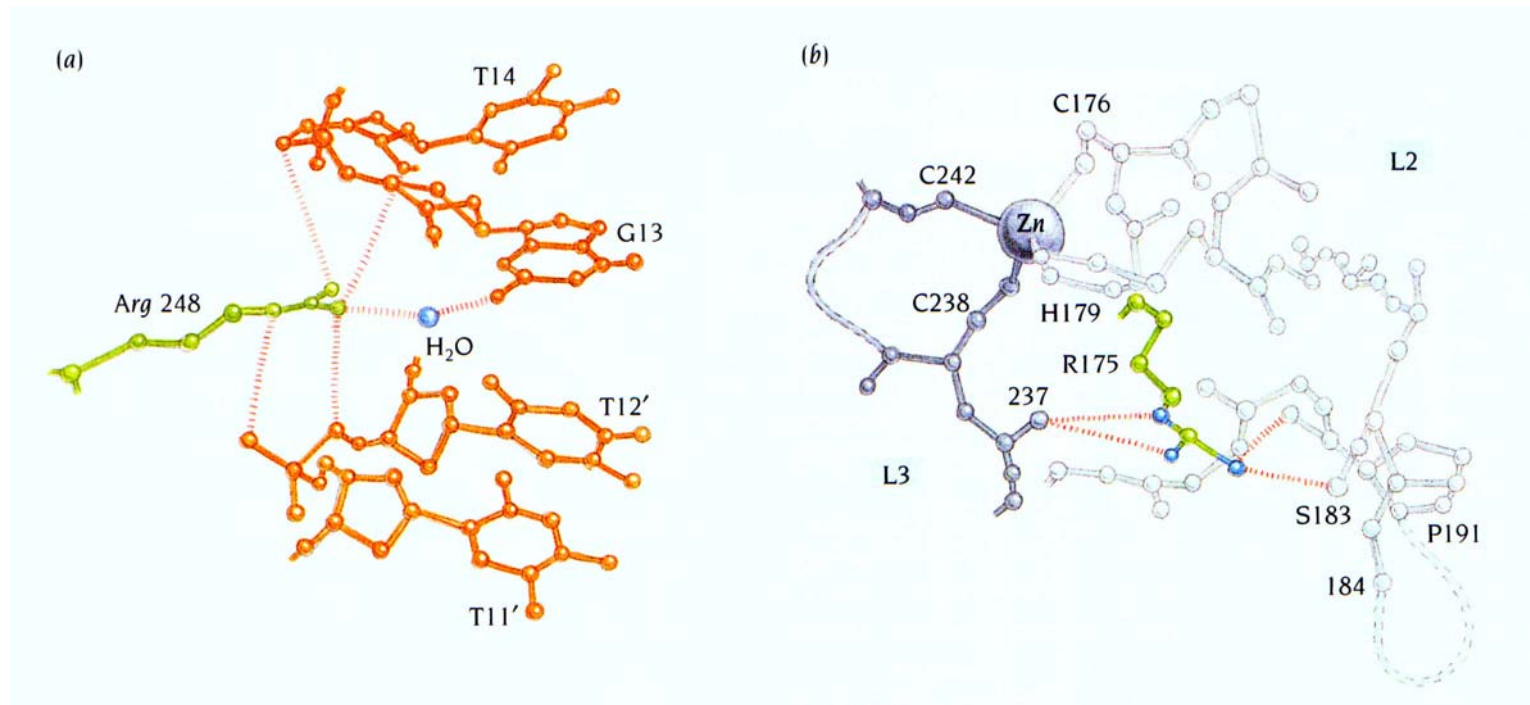
The nucleotide sequence of the 21-base pair DNA fragment was used for cocrystallization with the DNA-binding domain of p53. The p53 binds in a sequence-specific manner to the shaded region.

Two loop regions and one α helix of p53 bind to DNA



The interactions between DNA and p53 are sequence-specific. The C-terminal α helix and loop L1 of p53 bind in the major groove of the DNA. Arg 280 from the α helix and Lys 120 from L1 form important specific interactions with bases of the DNA. In addition, Arg 248 from loop L3 binds to the DNA in the minor groove.

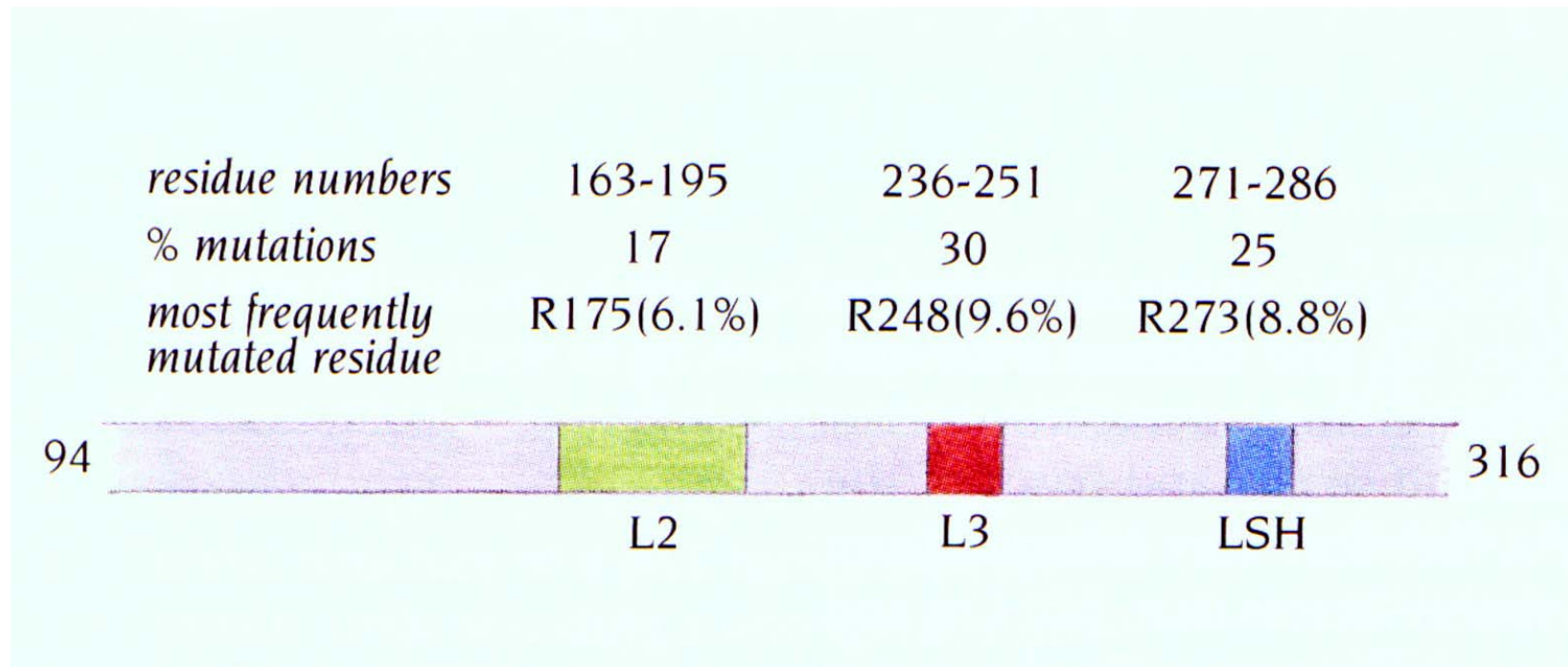
Two arginine residues of p53 are frequently mutated in tumors



(a) Interactions between Arg 248 and the DNA are shown. The side chain of Arg 248 is wedged into the minor groove and makes contacts with the sugar and phosphate groups of T12' and T14. In addition, a water molecule mediates a hydrogen bond between Arg 248 and the base G13.

(b) Interactions between L2 and L3 in p53. The zinc atom is bound to two cysteine residues in L3 and to one cysteine and one histidine in L2. In addition, residue Arg 175 from L2 forms two hydrogen bonds to the main chain C=O group of residues 237 in L3 and one hydrogen bond to the side chain of Ser 183 in L2. Mutations in residues 175, 183 or 237 would distort loop L3 and prevent proper binding of p53 to DNA.

Tumorigenic mutations occur mainly in three regions involved in DNA binding



These regions are loops L2, L3 and a region called LSH which comprises part of β strand 9 as well as the C-terminal α helix.