Part 4

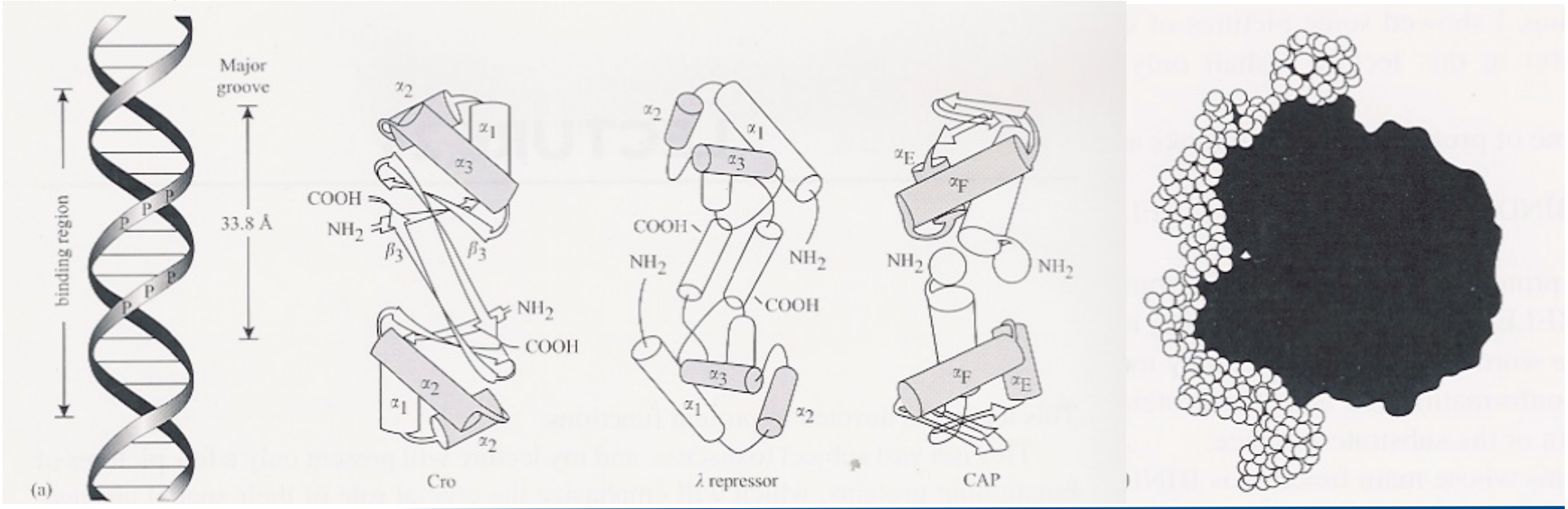
- Protein-DNA interactions
- Protein-RNA interactions
- Protein-ligand interactions

- Structural bioinformatics :
  - Comparison between protein structures
  - Classification of protein structures

Protein-DNA interactions

Protein function: Binding → Release → Transformation
Some proteins perform only some of these actions.
- Binding / Release of various molecules (protein, DNA, ligand, ...)
- Transformation: Chemical transformation, conformational change (of the protein and/or the substrate), movement of the protein or the substrate in space

Example of binding/release: protein-DNA interactions
To bind to DNA: a portion of the protein surface must be approximately complementary to the surface of the double helix (taking into account the residual flexibility)

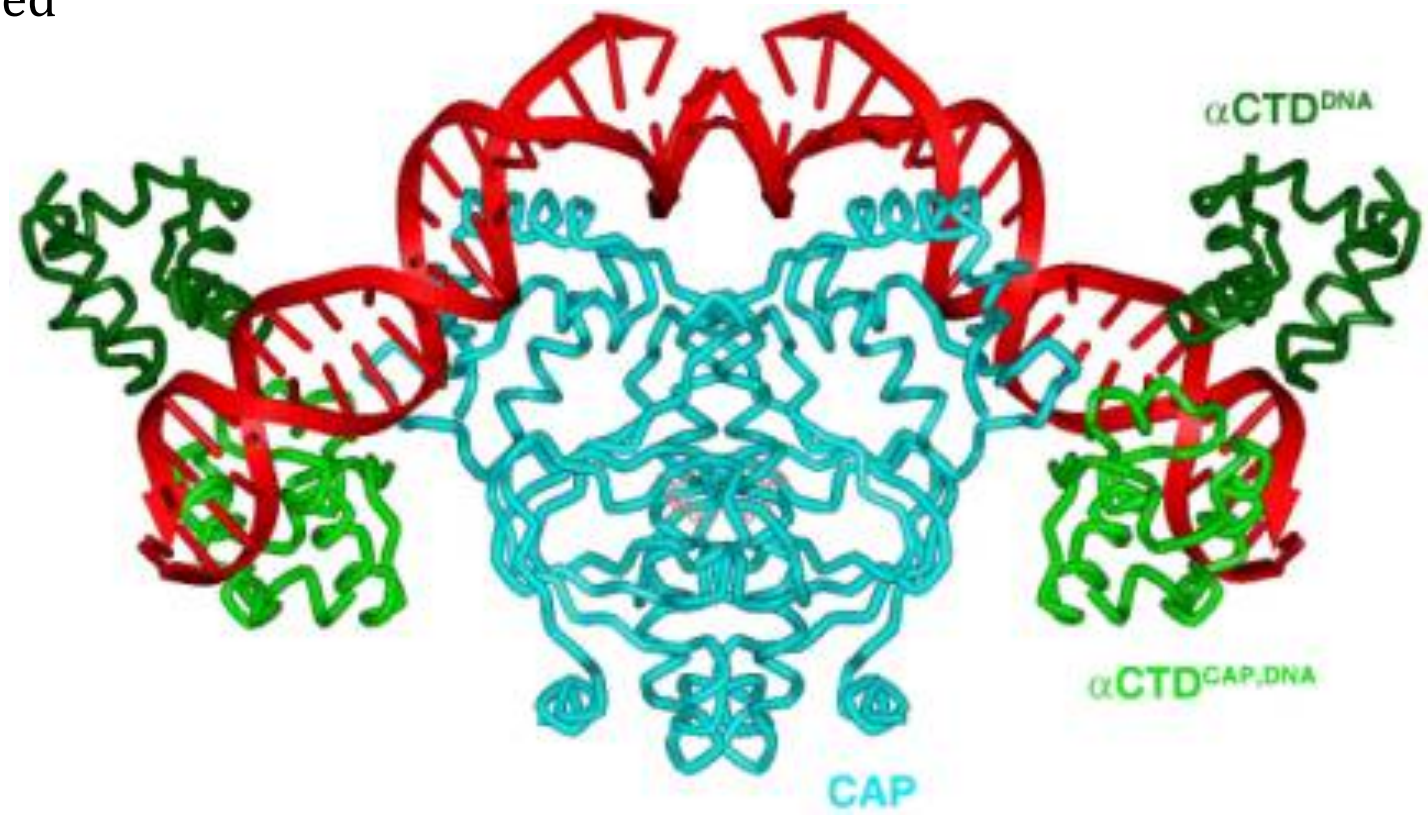


Principles of protein-DNA recognition

The information encoded in the DNA must be:

- replicated
- read and transcribed/translated into gene products (RNA, proteins) depending on the cell type, stage of development, cell cycle stage, as a response to an external stimulus
- wrapped, packaged
- repaired, ...

these processes require many proteins - specific or nonspecific.



Protein-DNA interactions

When the complementarity of the surfaces is good: the side chains of the protein penetrate deeply into the DNA → specific recognition (depends on the protein and DNA sequences)

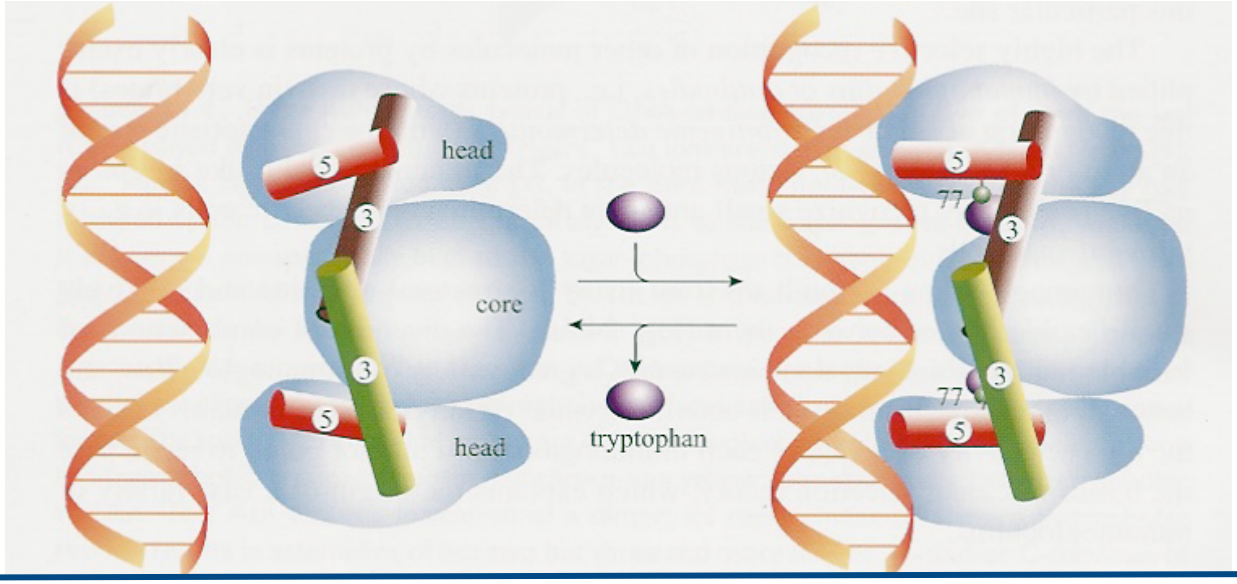
Protein + DNA ↔ Protein-DNA complex

Affinity : $K_{binding}$ = [protein-DNA] / [protein] [DNA]

Specificity = $K_{binding}$ (target) / K_{bindng} (non-target)

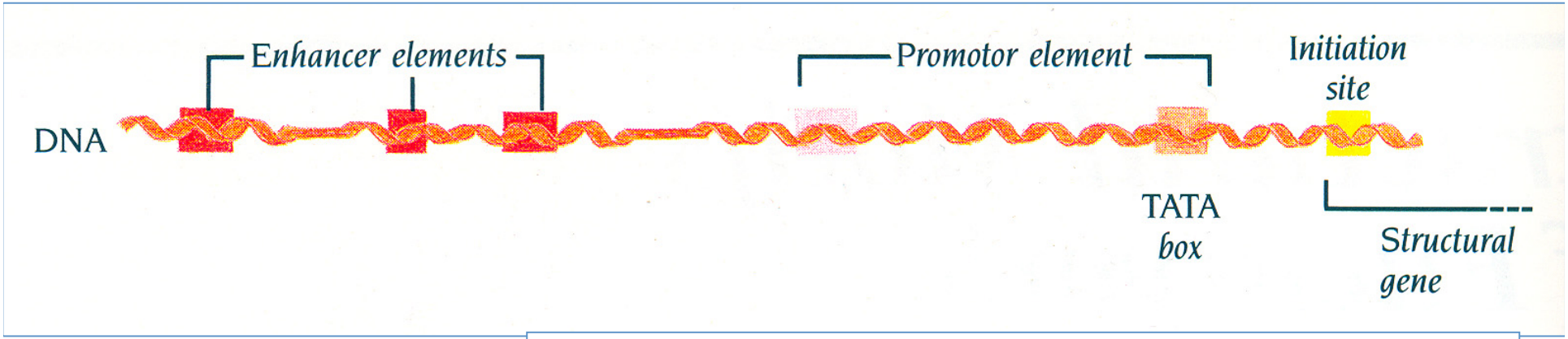
Sometimes, the DNA-protein binding requires a co-factor, which changes or distorts the protein structure: inactive state → active state

For example, the Trp-repressor (inhibits the synthesis of proteins required for the synthesis of Trp) - allosteric stimulation because the Trp binds to a site that is different from the one that binds DNA - here helix-turn-helix (HTH) motif (see later)

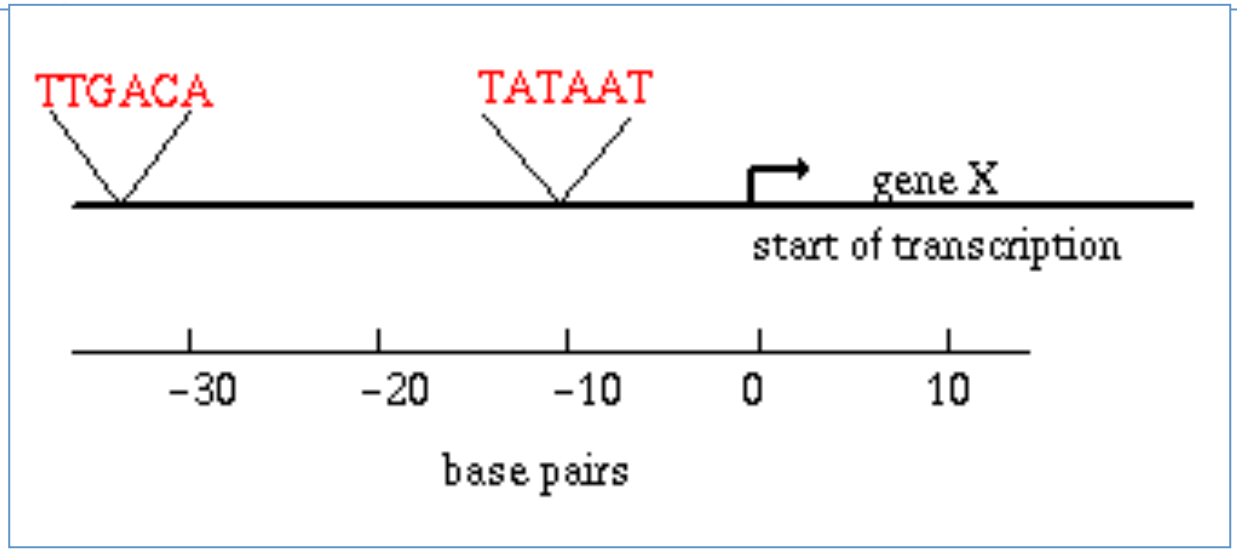


Protein-DNA interactions

Transcriptional control in eukaryotes : more complex and less well understood than that of prokaryotes, which is simply of the type on / off



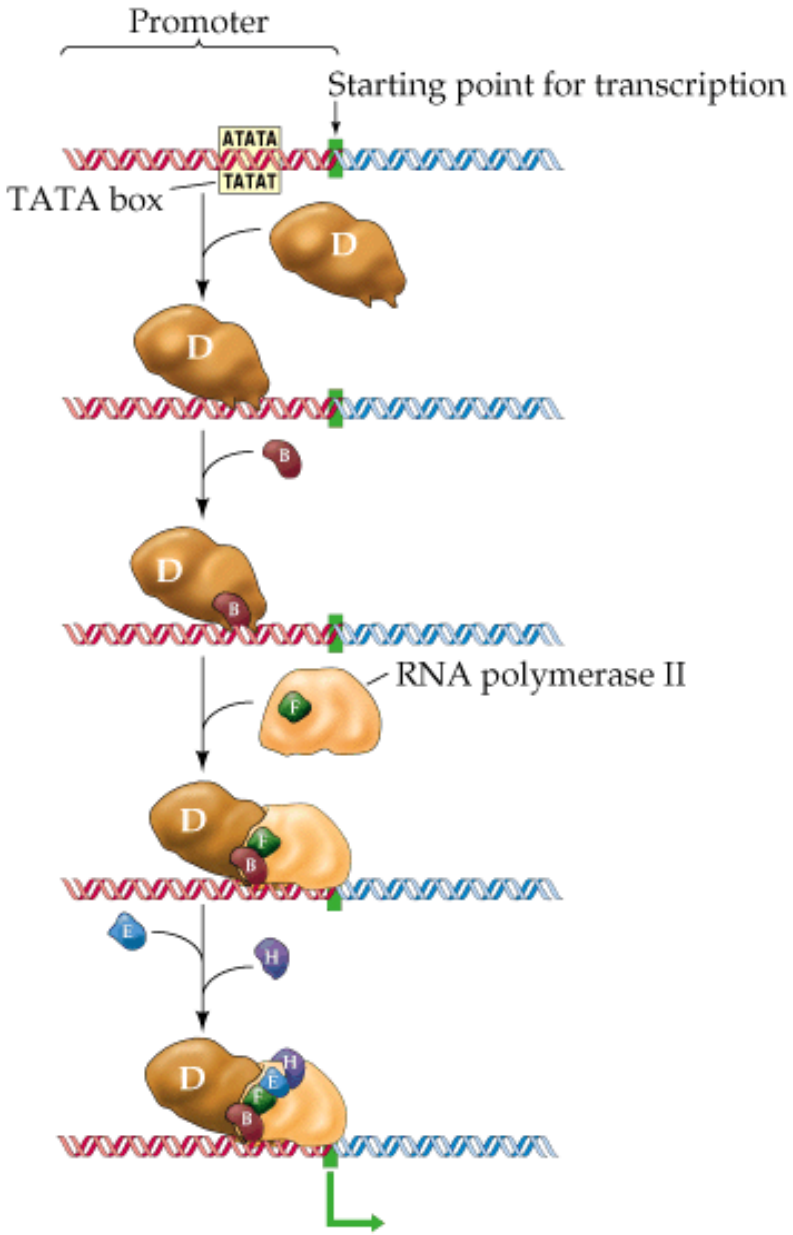
The TATA box is the general promoter sequence in eukaryotes. It is recognized by TBP (TATA box binding protein)



Protein-DNA interactions

TBP is the central component of a protein complex, which binds a set of factors such as TFIIA, TFIIB, TFIIE and finally the RNA polymerase II to form the *preinitiation complex* (PIC).

TBP is the classic example of a protein that recognizes DNA by its characteristic deformability produced by an AT-rich sequence, rather than by "reading" directly the base pairs (recognition of structure rather than sequence)

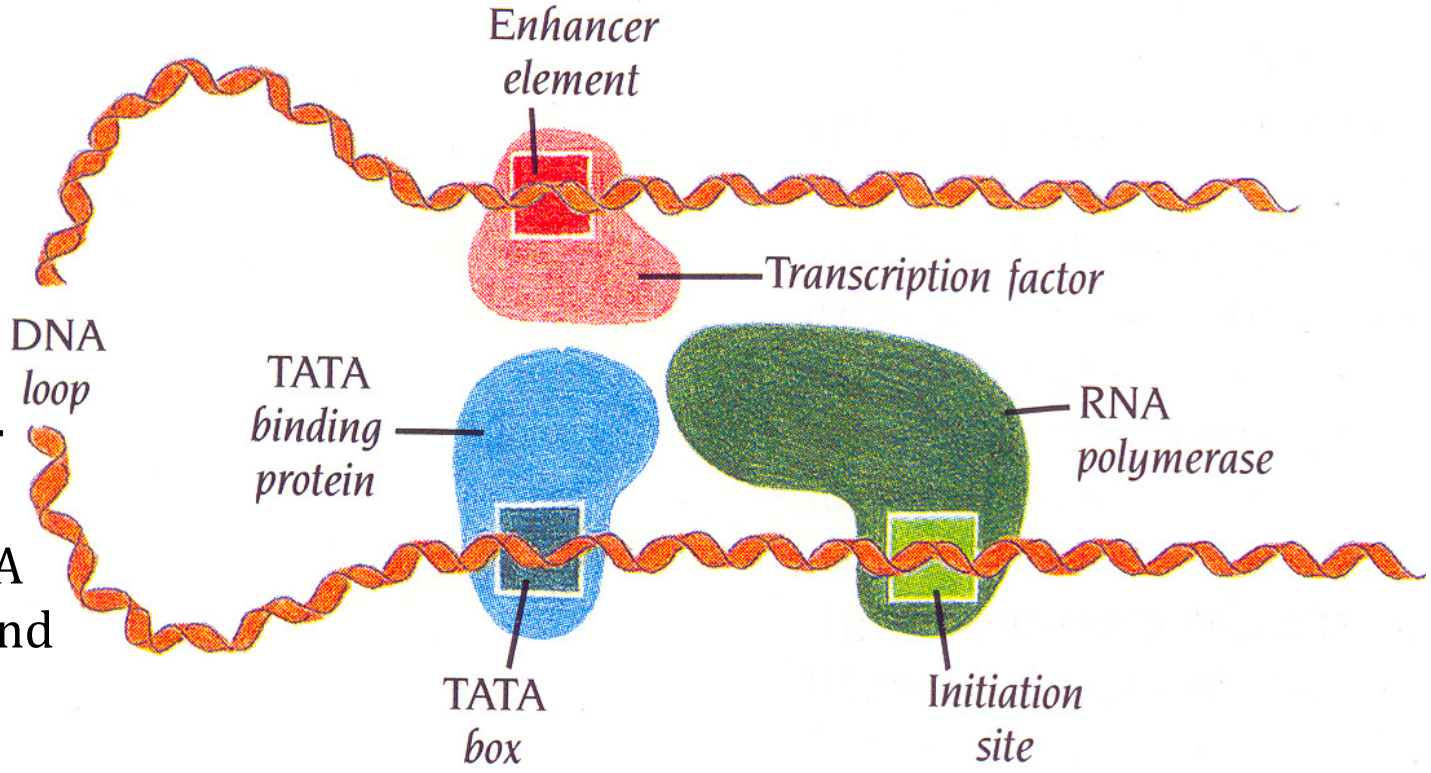


Protein-DNA interactions

In general, a transcription complex may contain four types of proteins:

- Basal factors, e.g. TBP
- Activators that bind to sites sometimes distant along the DNA (enhancers)
- Repressors that bind to certain DNA sites to prevent the binding of activators
  -> makes the gene silent
- Coactivator that bind to basal factors and activators

One or more regulatory proteins can bind to DNA sometimes far upstream of the initiation site and interact with TBP or RNA polymerase -> curvature of the DNA loop between TBP and other transcription factors



Protein-DNA interactions

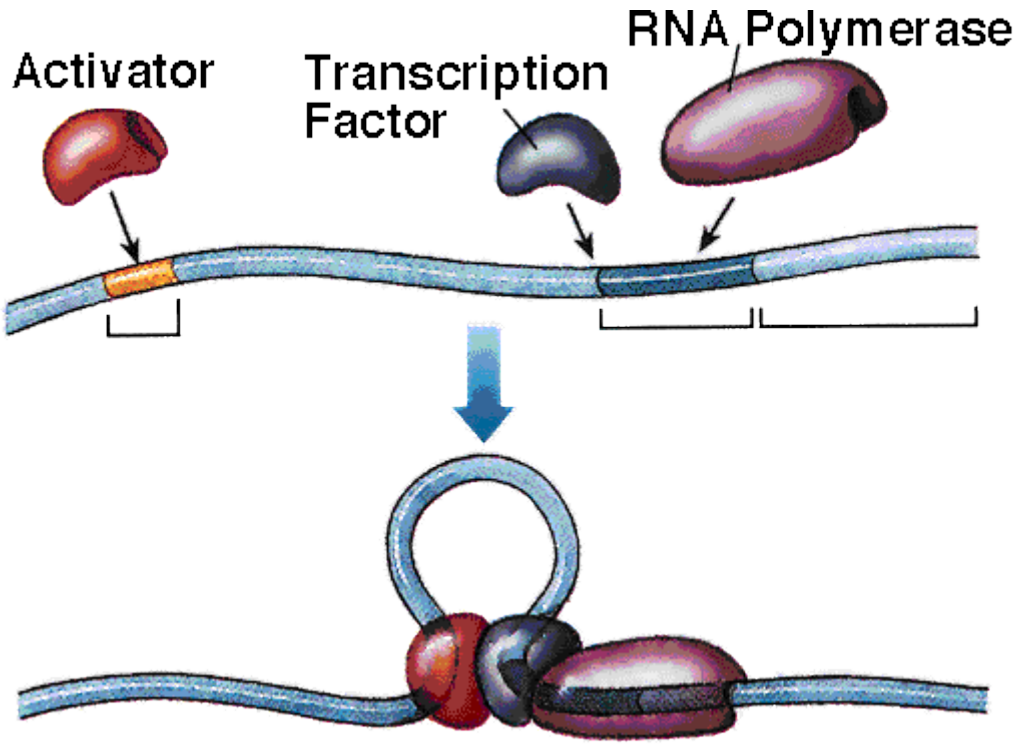
The DNA binding to transcription factors confers specificity to the gene to be transcribed - spatial specificity and temporal specificity

Activator proteins bind to specific DNA sequences.

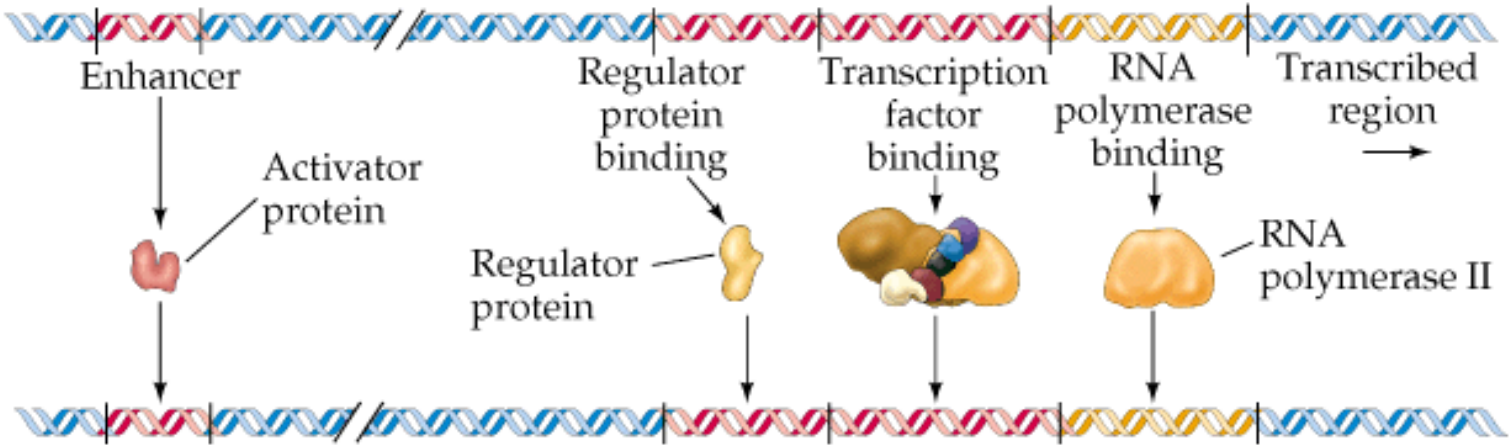
The specificity of an activator is determined by the DNA sequence to which it binds.

Activators bind cooperatively. They contain portions of their surface for DNA binding, and one or more portions of their surface for binding to other transcription factors – this increases specificity.

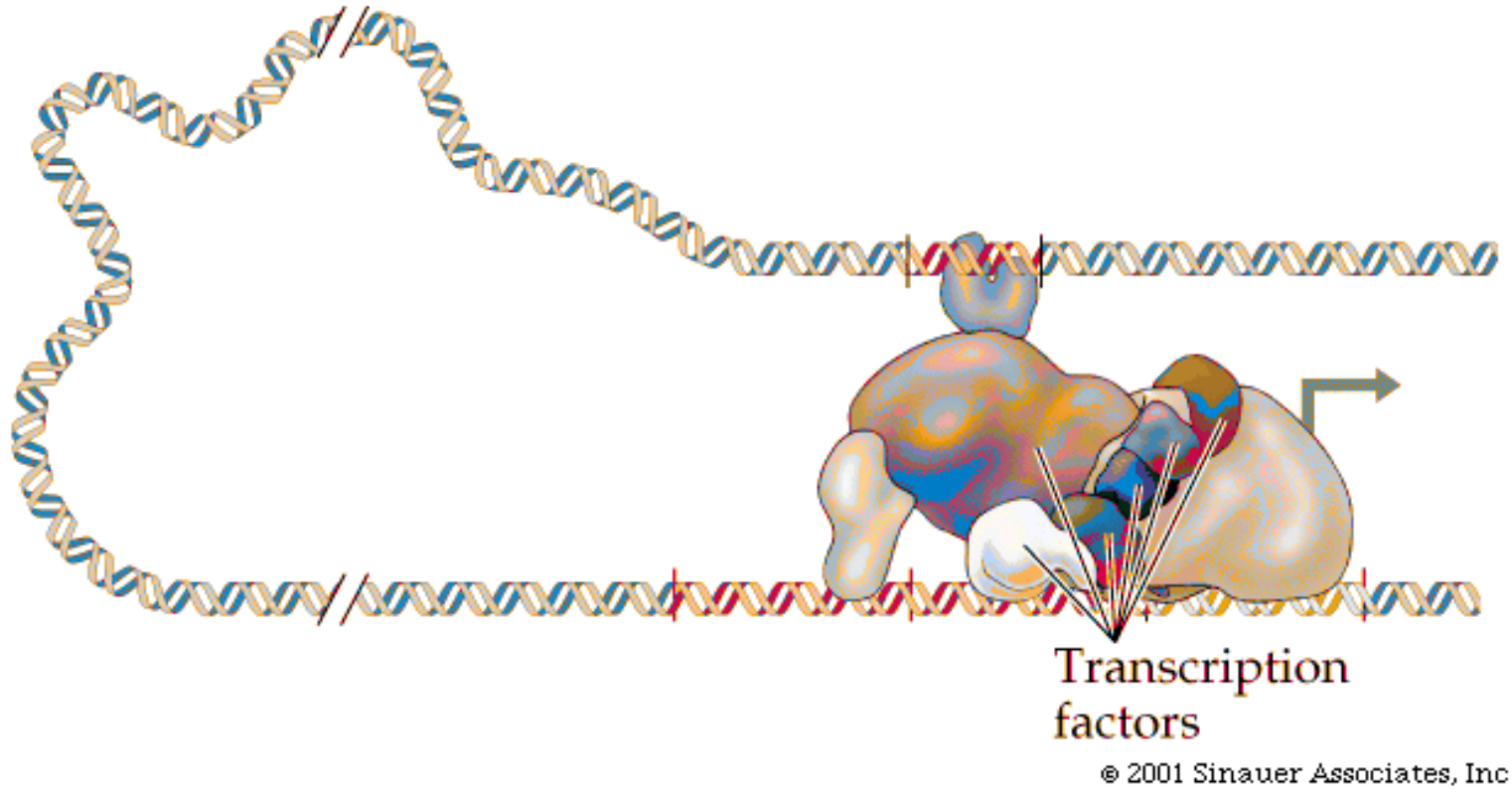
Activators have binding motifs, usually to the major DNA groove, of the type: HTH, homeodomain, zinc finger, leucine zipper (see later)



Protein-DNA interactions



Importance of quaternary complexes for function



Protein-DNA interactions

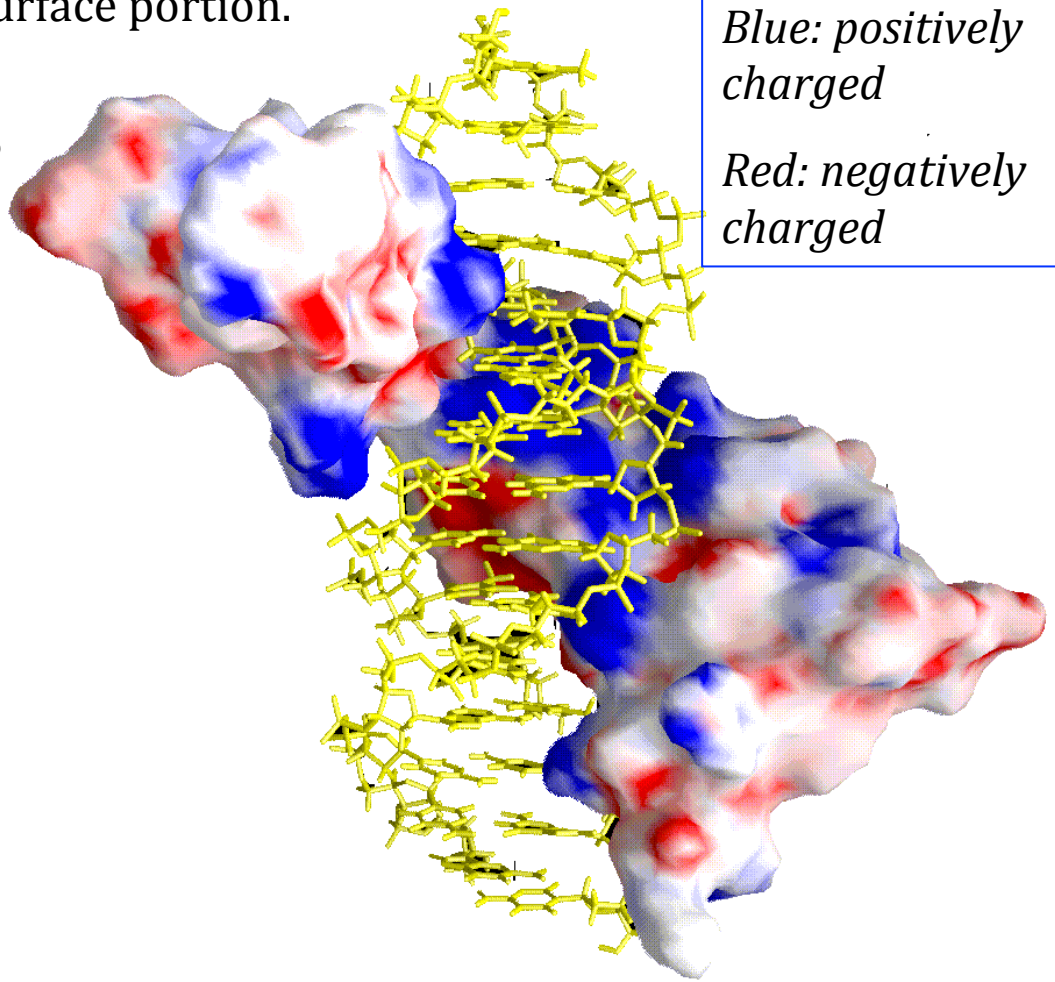
Electrostatic forces are felt at relatively long distance, but they are not very specific.

The surface of the DNA is negatively charged due to the phosphate groups => attract proteins having a positively charged surface portion.

When the protein and DNA are close, owing to the electrostatic interactions, other forces that are shorter range become effective.

These are mainly :

- H-bonds between amino amino acids and nucleobases,
- cation-π interactions between positively charged amino acids and nucleobases,
- amino-π interactions between amino acids bearing a partially charged group and nucleobases
- π-π interactions between aromatic amino acids and nucleobases



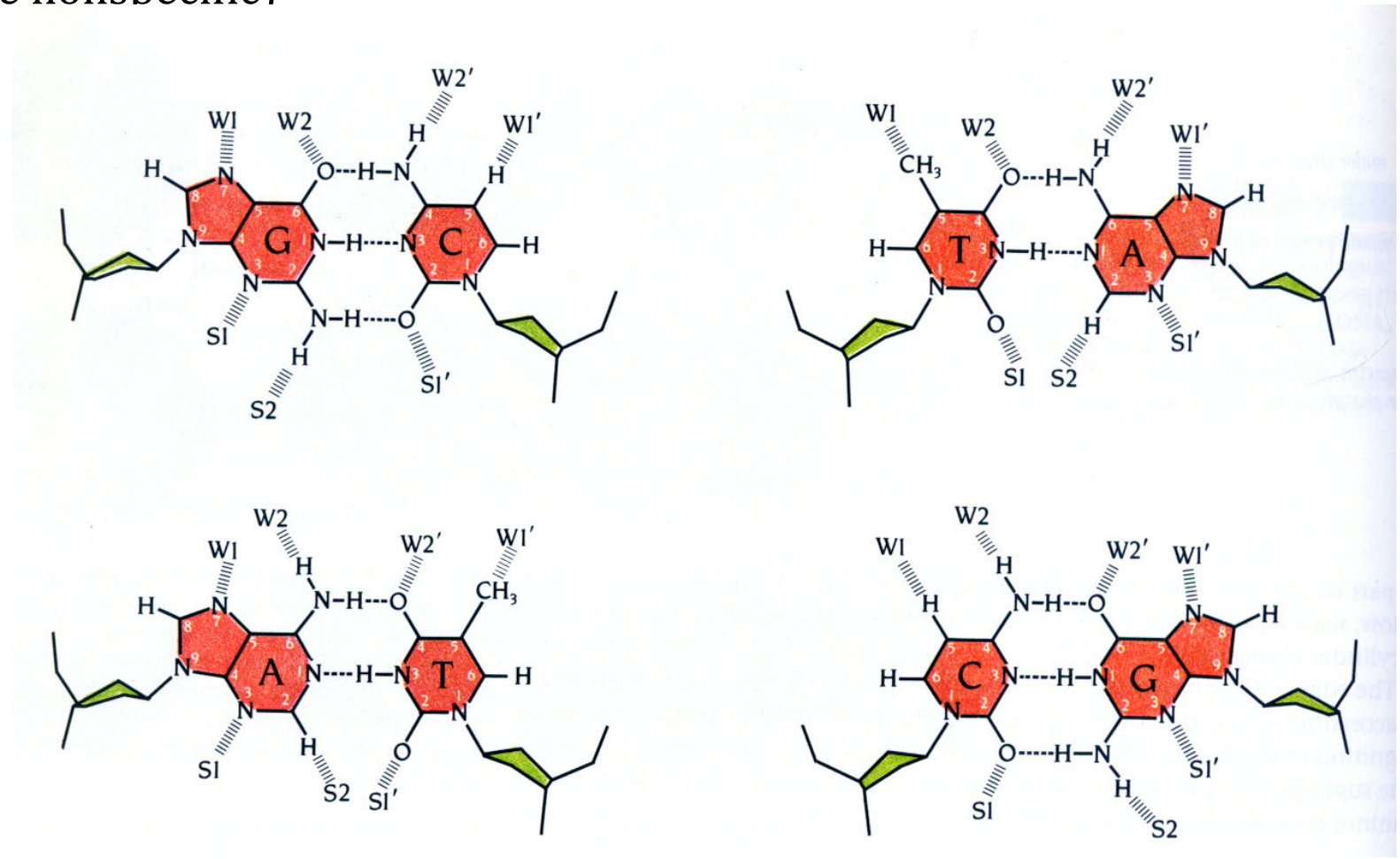
Blue: positively charged

Red: negatively charged

Protein-DNA interactions : specific binding

The only regions of the DNA that are accessible to a specific interactions are inside/at the bottom of the grooves (interactions with the sugar-phosphate backbone are nonspecific)

At the bottom of the grooves, the atoms H, N and O can form H-bonds with the amino acids in the protein.

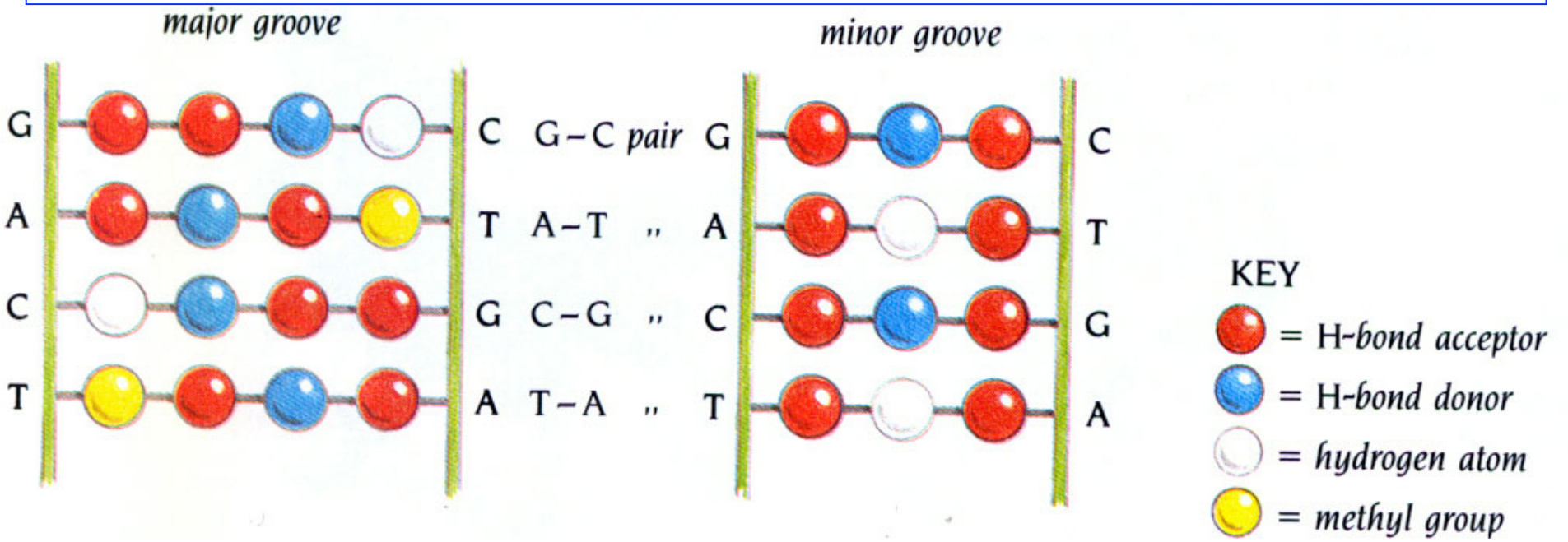


Protein-DNA interactions : specific binding

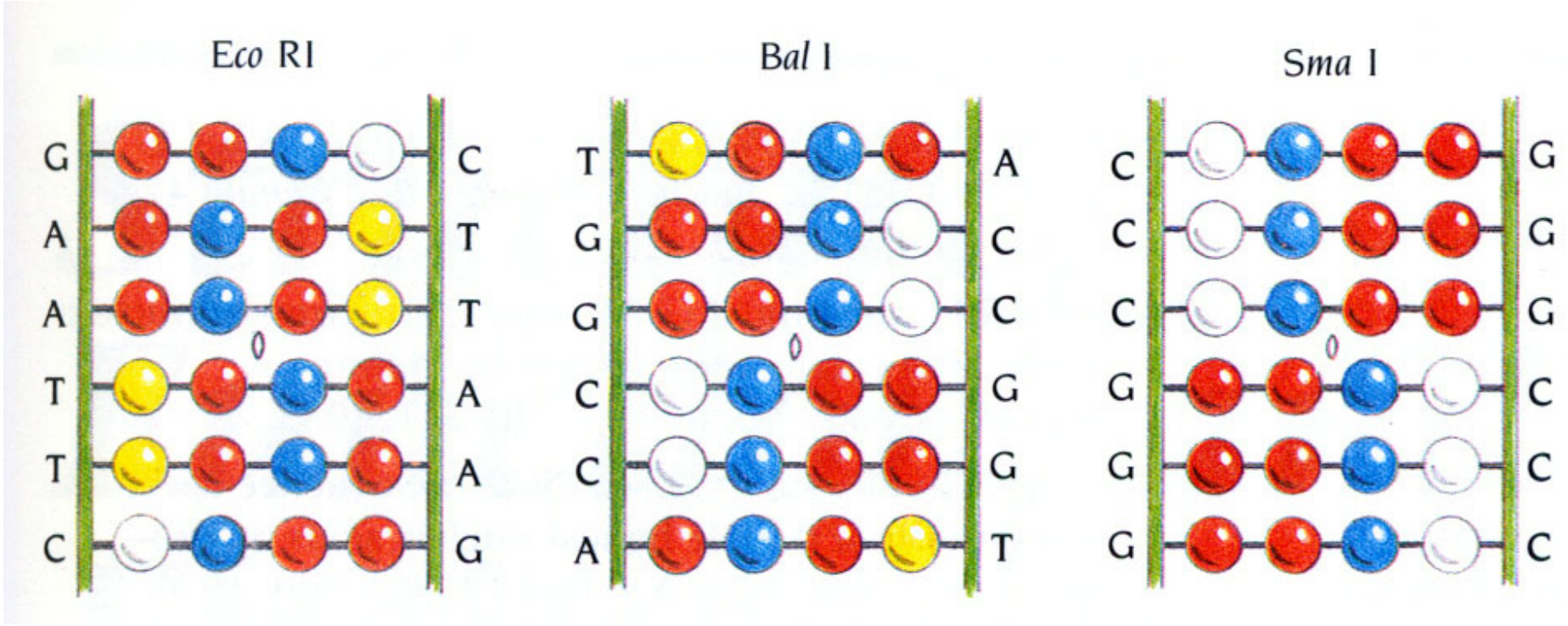
More opportunities to bind in the major groove than in the minor groove
=> larger specificity. Because:

- Major groove is wider, and the bases are more accessible
- More possible H-bonds with different amino acids

Specificity of protein-DNA binding is important because it determines which genes are transcribed
(note: there are certain proteins that bind DNA nonspecifically, e.g. histones, which compact DNA).



Protein-DNA interactions : specific binding



For example, using the colour code of the previous slide:

Recognition of a pattern of 6 bases
for three restriction enzymes: Eco RI, Bal I and Sma I.

=> The motifs that are recognized are very different => specificity

Protein-DNA interactions : specific binding

There is no 1:1 correspondence between bases and amino acids, but there are marked preferences.

For example, in a set of 129 protein-DNA structures, these are:

	Gua	Cyt	Ade	Thy	Total
Arg (R)	98	8	19	24	149
Lys (K)	30	6	4	9	49
Ser (S)	12	2	1	3	18
Asn (N)	7	10	18	7	42
Gln (Q)	6	2	16	2	26
Glu (E)	1	10	1	0	12
Total	154	38	59	45	296

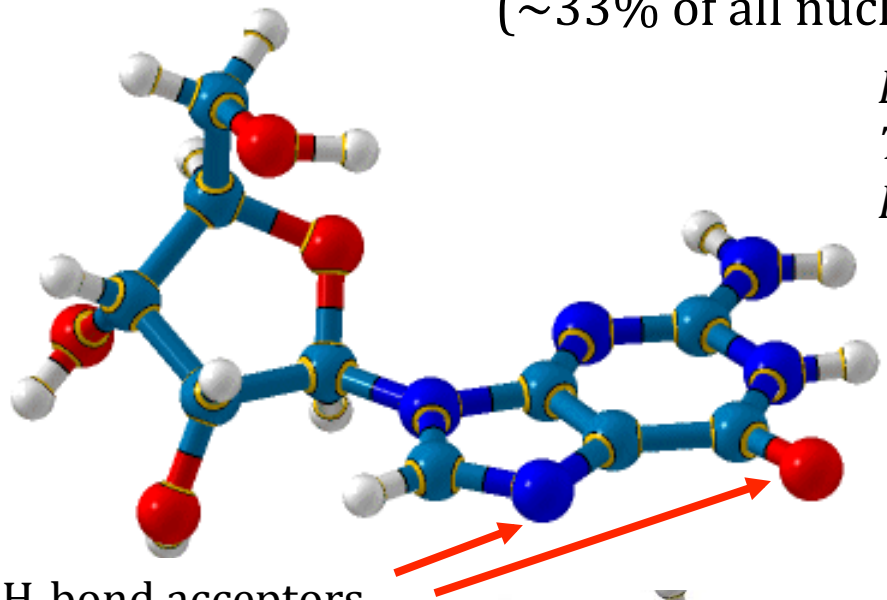
Marked preferences: Arg – Gua Lys – Gua Asn/Gln – Ade

…..

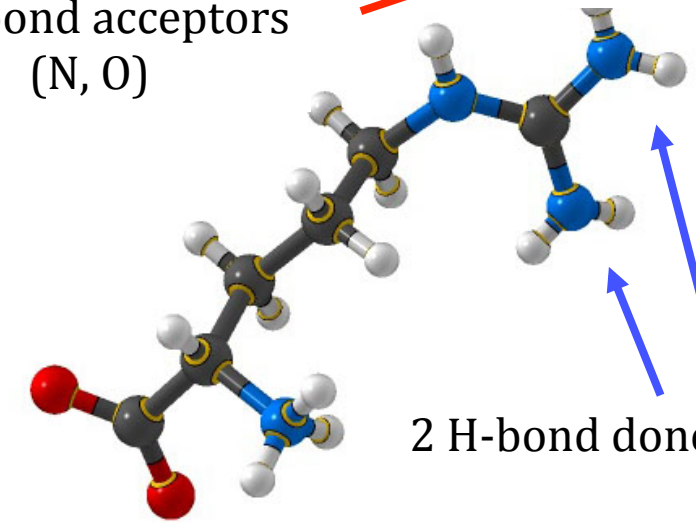
Why?

Protein-DNA interactions : specific binding

Arg – Gua : perfect H-bond association !! Double H-bond
(~33% of all nucleobase-amino acid interactions)

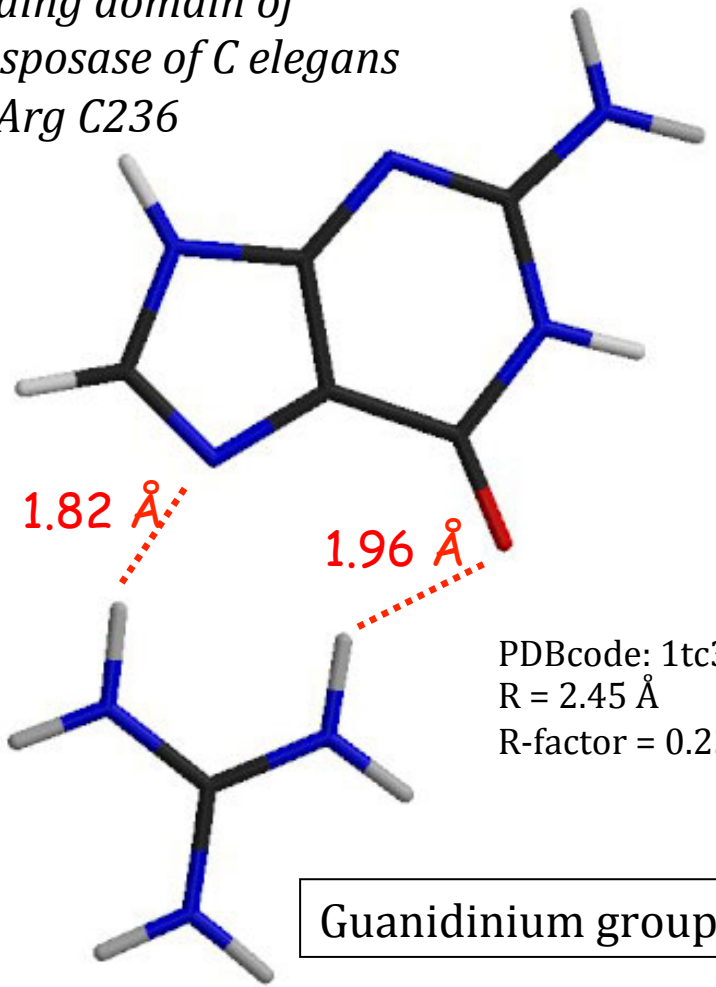


2 H-bond acceptors
(N, O)



2 H-bond donors (H)

DNA binding domain of
Tc3 transposase of C elegans
Residue Arg C236

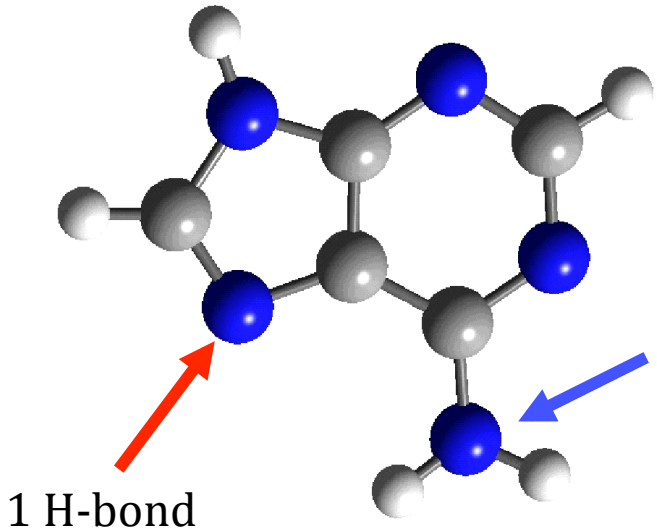


PDBcode: 1tc3
R = 2.45 Å
R-factor = 0.234

Guanidinium group

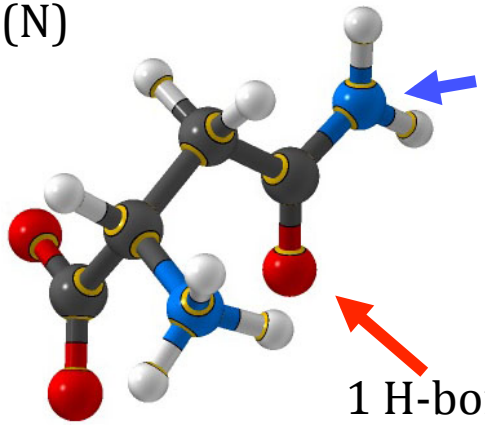
Protein-DNA interactions : specific binding

Asn/Gln – Ade : another frequent H-bond association
(~11% of all nucleobase-amino acid interactions)



1 H-bond acceptor (N)

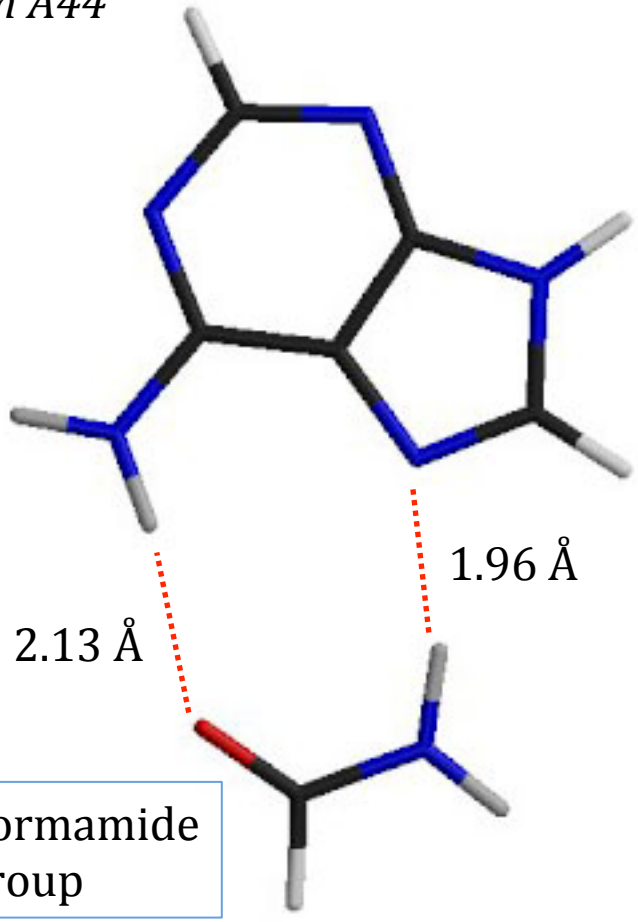
1 H-bond donor (H)



1 H-bond donor (H)

1 H-bond acceptor (O)

Pit-1 Pou domain
Residue : Asn A44



1.96 Å

2.13 Å

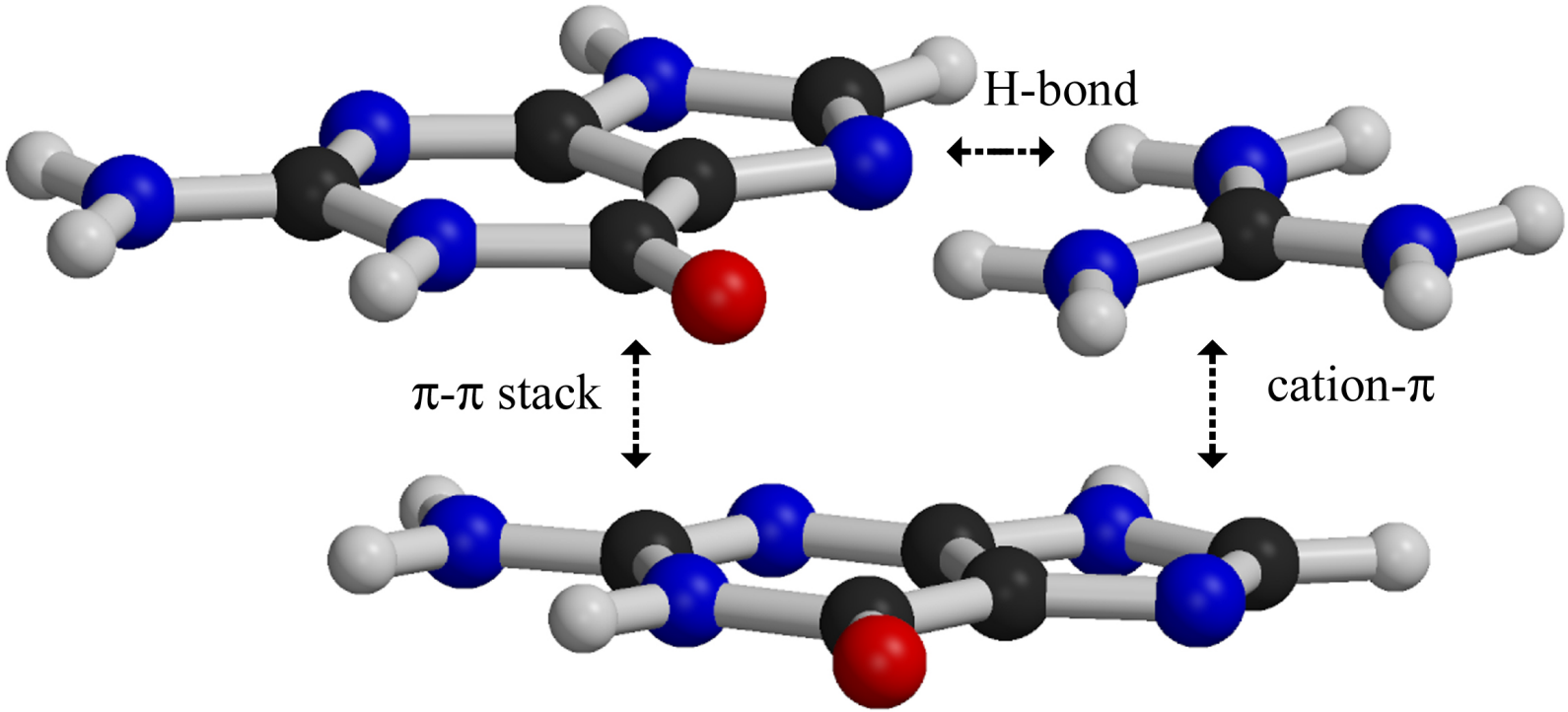
Formamide group

Protein-DNA interactions : specific binding

Another frequent interaction:

Cation-π/H-bond stair motif - involves two nucleobases and 1 amino acid.

Three types of interactions: π–π stacking, H-bond and cation-π.

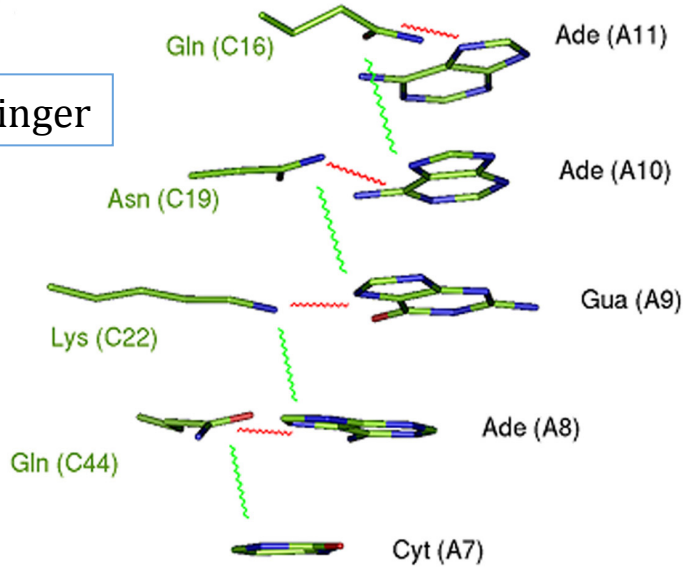


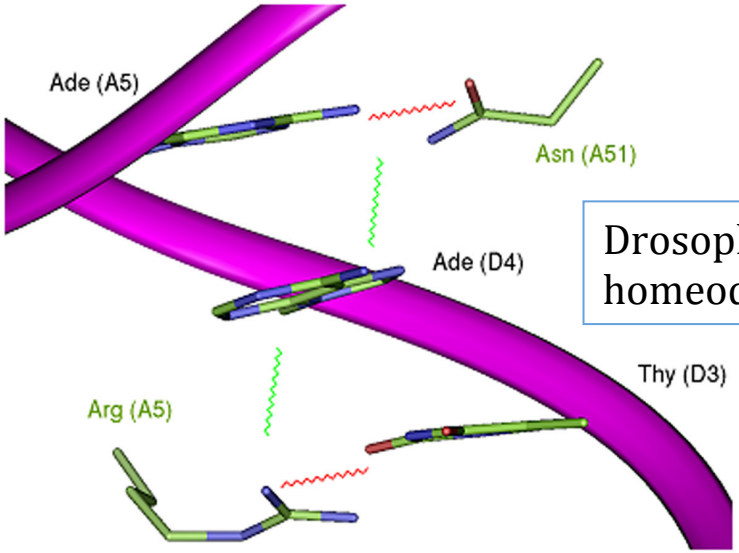
DNA binding domain of Tc3 transposase of *C elegans*

Residue : Arg C236 ; Bases Gua A7 et A8

Protein-DNA interactions - Typical stair motifs

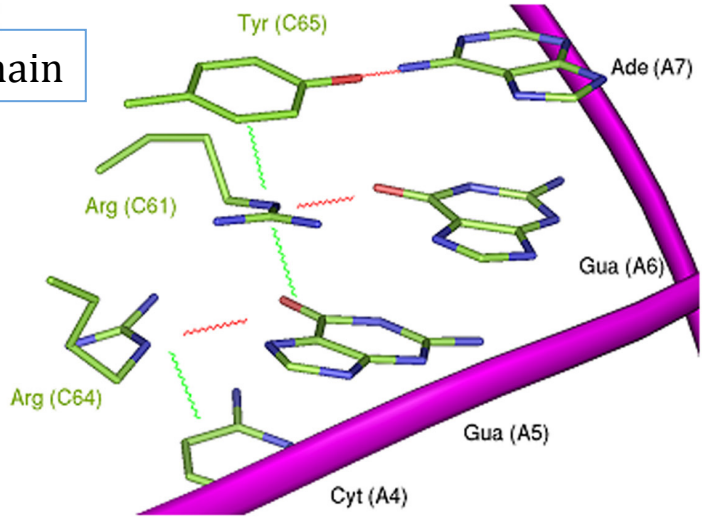
Zinc finger

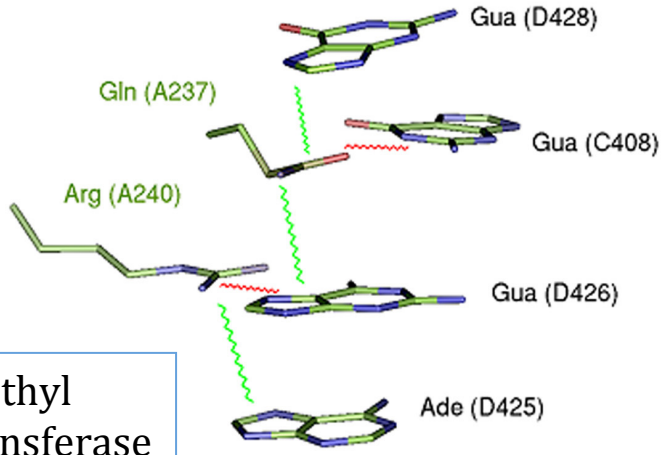




Drosophila homeodomain

Ets domain





Methyl transferase

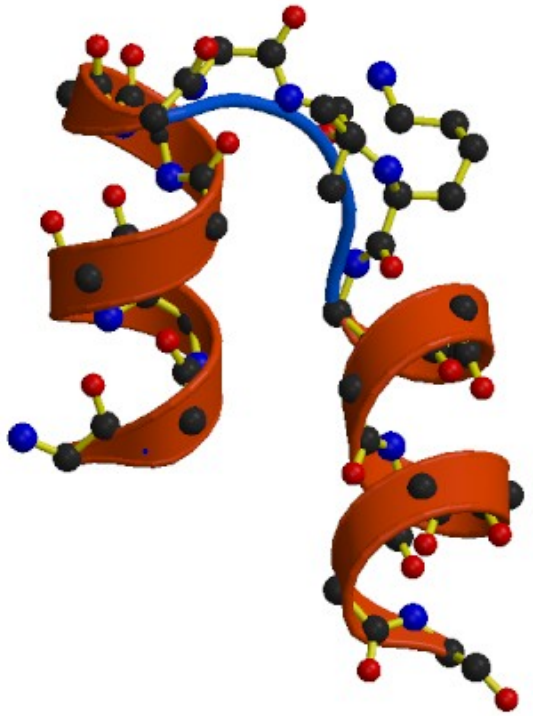
Classification of protein-DNA motifs

1. HTH type domains

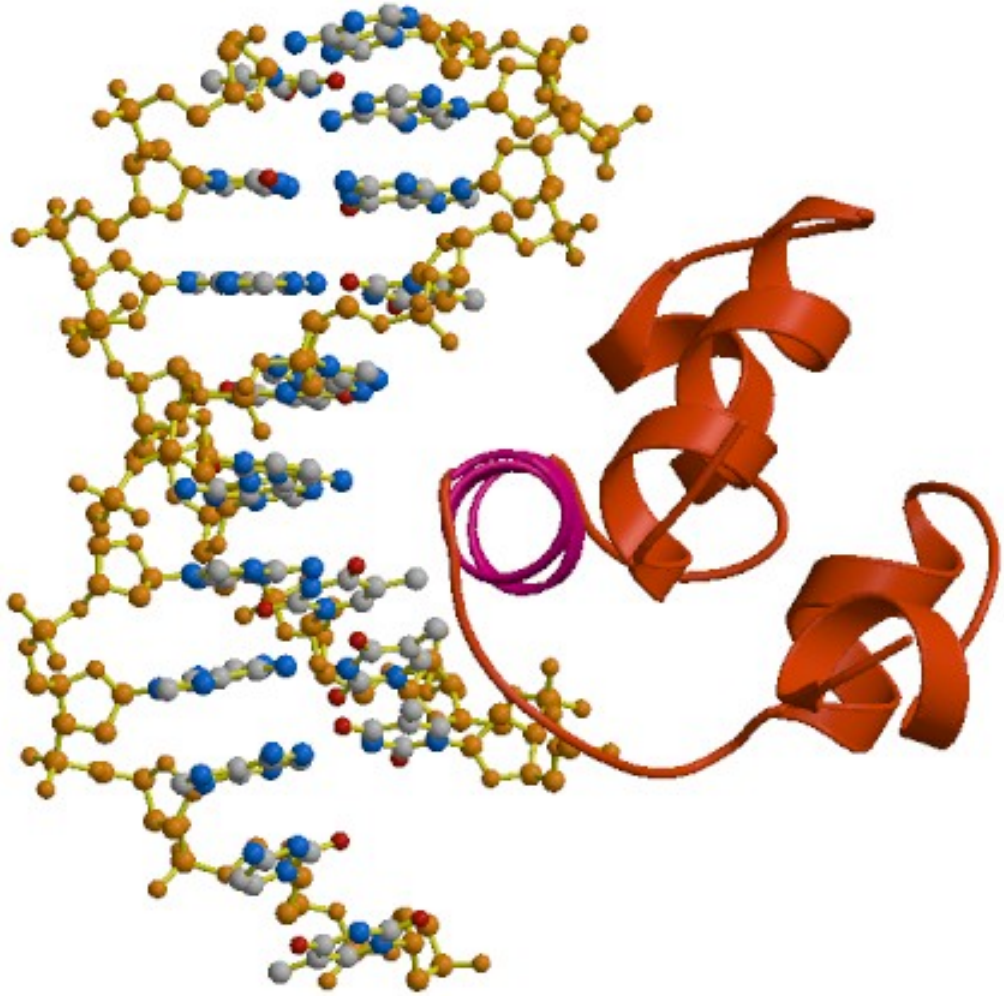
The HTH structure was the first DNA binding motif that was discovered – it is found in protein domains of various structure, origin and function, such as regulatory proteins of transciption in prokaryotes, homeodomains that are involved in cell differentiation during development of eukaryotes, and histones whose role is to package DNA.

Independently of its origin, the HTH motif has a characteristic shape:

- 2 helices forming an angle of about120 °,
- often connected by a specific turn of 3 residues with Gly in position 1, whose dihedral angle ϕ is positive.



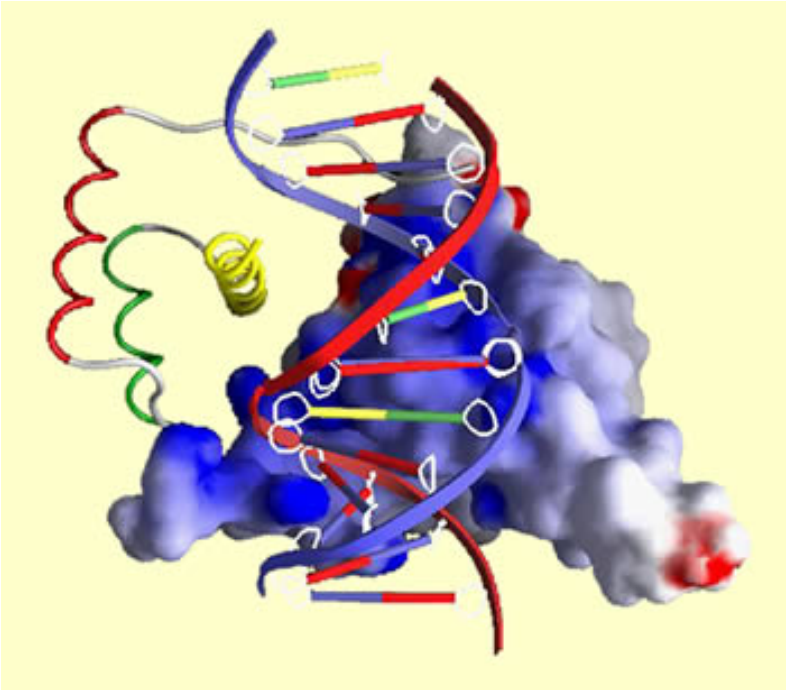
The second helix of the motif is the recognition helix, which protrudes from the protein surface and enters into the major groove of the DNA => specific protein-DNA contacts.

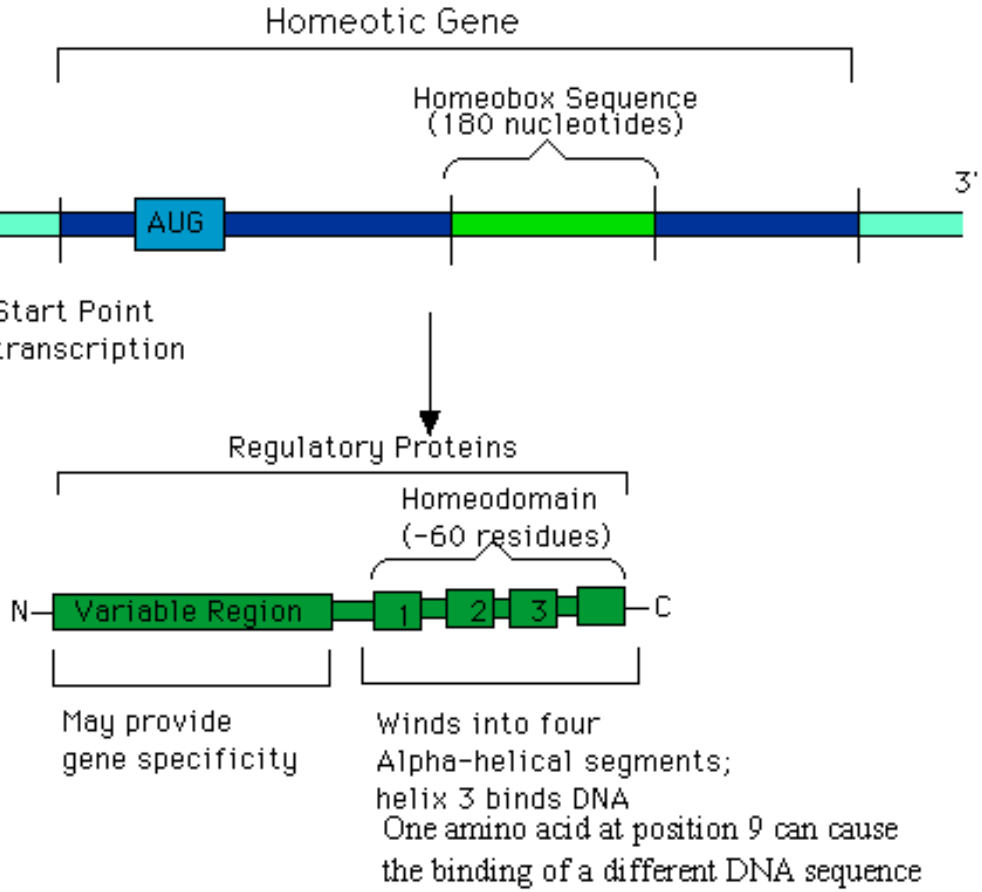


In addition to the canonical HTH motif, these domains contain a 3rd helix of variable orientation, sometimes a 4th helix, a small β-sheet or a flexible N-or C-terminal region => confer HTH domains their structural and DNA binding specificities : they often touch the edge of the major groove penetrated by the recognition helix, or the adjacent minor groove.

In addition, they bind as monomers or dimers -> increased specificity in binding itself + constraints associated with protein-protein interface.

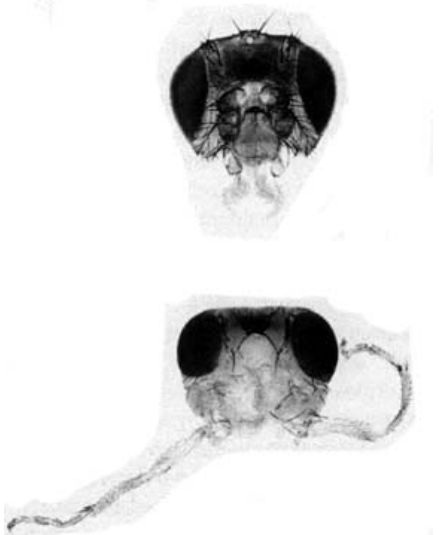
Example of HTH motif: homeodomains, which are proteins of about 60 residues with a flexible N-terminal tail + 3 helices, of which the last two form the HTH motif. *Function*: DNA binding domains of transcription factors.

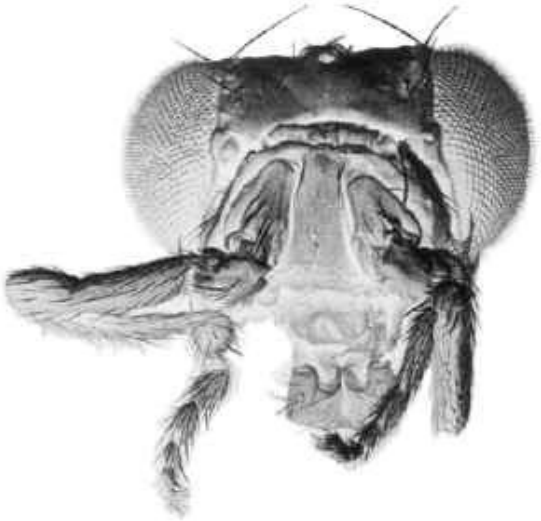




Homeodomains are encoded in homeoboxes, which are part of homeotic genes - their variable part confers a functional specificity

Homeotic genes code for regulatory proteins that constitute a major control network.
They were first discovered in animals, especially Drosophila, where it was found that mutations in these genes were able to convert some full part of the body into another.
For example, one of these mutations leads to the growing of a leg instead of an antenna (the antennapedia mutation).



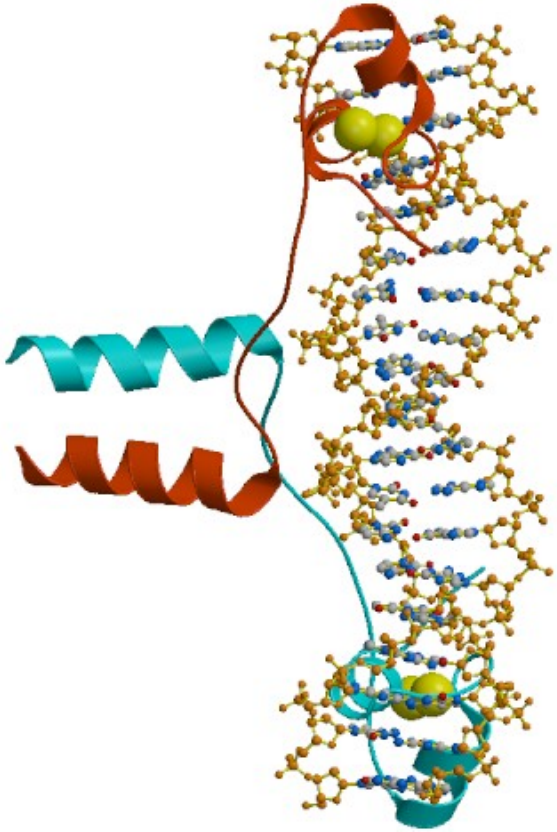


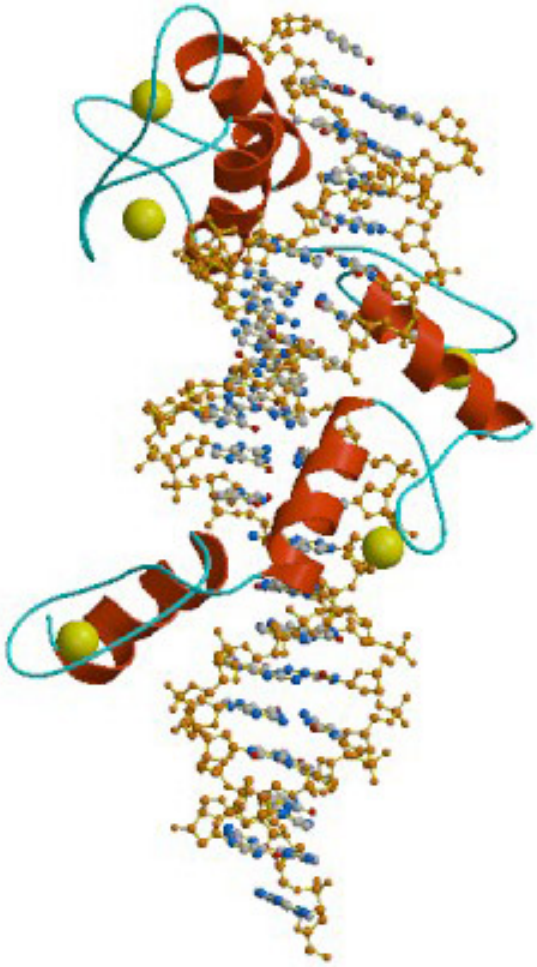
More than 20 homeotic genes were detected by mutations in Drosophila. They control major embryonic processes such as dorsal/ventral and anterior/posterior differentiation, the formation of eyes, legs, and wings.

Classification of protein-DNA motifs

2. Zinc fingers

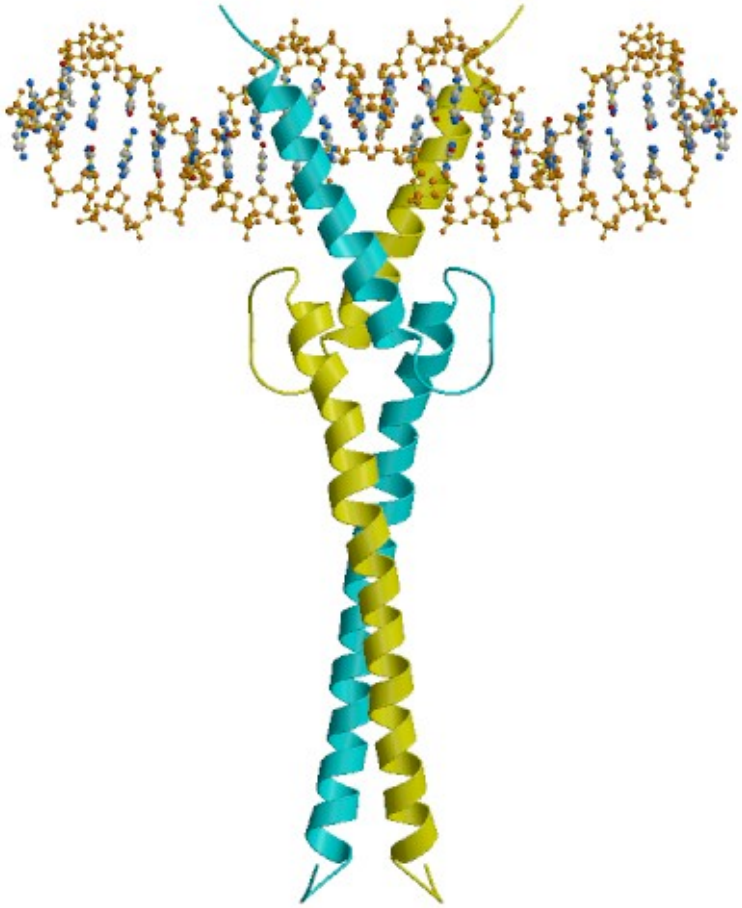
Motifs containing Zn: 3 classes of different structures, where Zn is linked by His and/or Cys

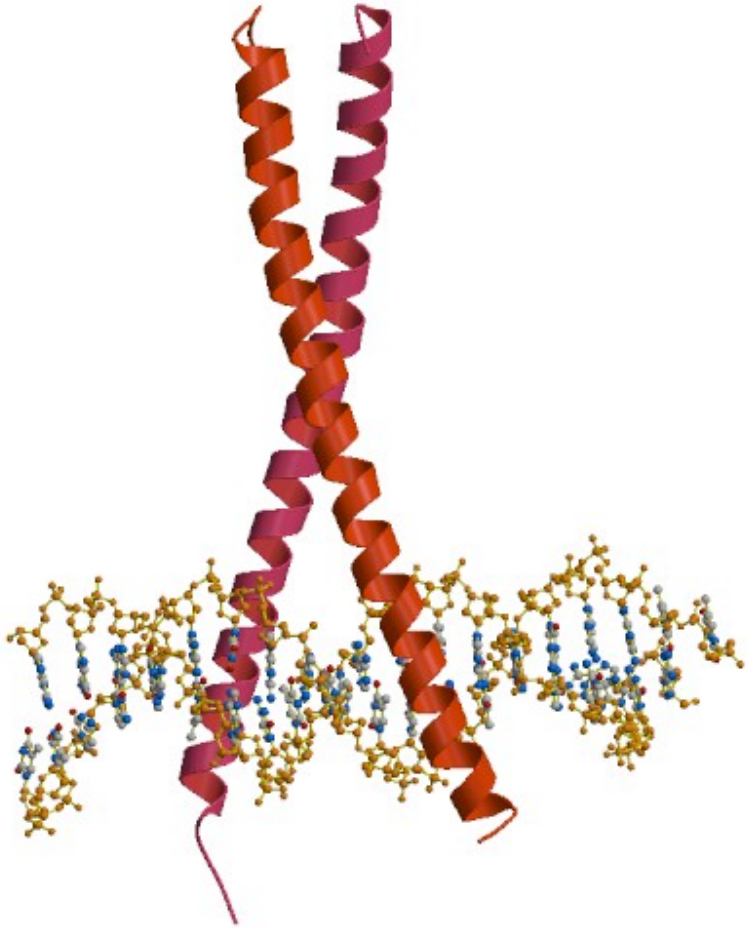




Classification of protein-DNA motifs

3. Leucine zippers

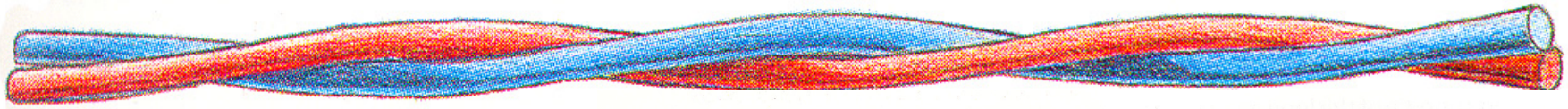




N-terminal part is basic and positively charged - interacts with DNA - is structured upon contact with DNA.

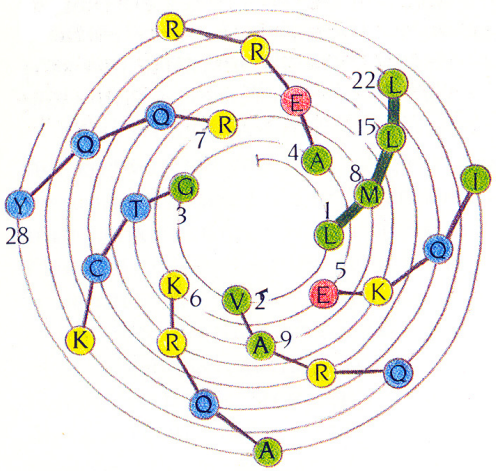
The helices of the two monomers enter into the major groove of either side of the DNA – they hold it as a pincer.

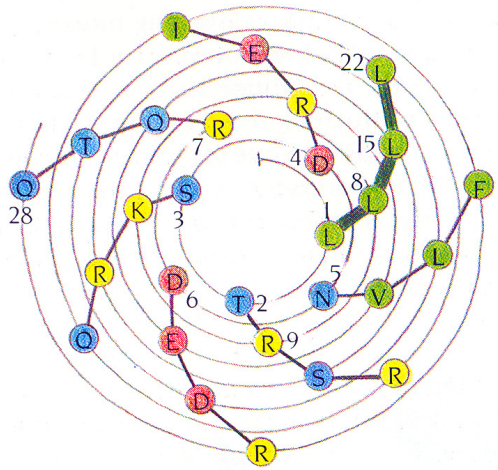
=> Recognize palindromic sequences of DNA in two successive major grooves.

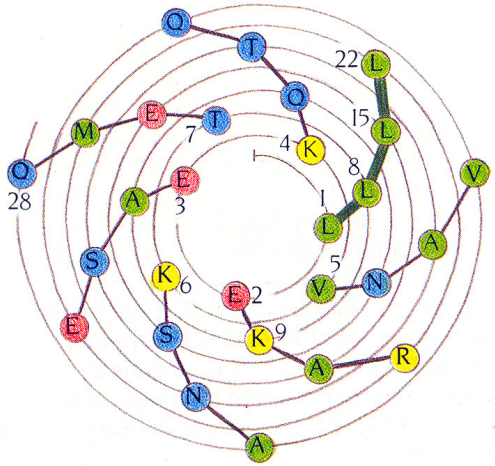


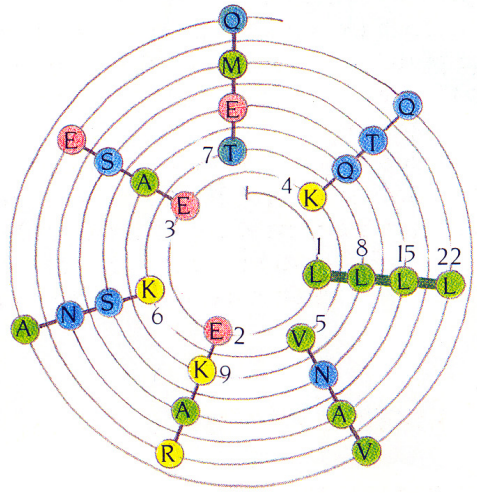
C-terminal region: coiled-coil structure = two helices wound around each other, with the inner faces hydrophobic - typically Leu

Another example of coiled-coil structure: keratin

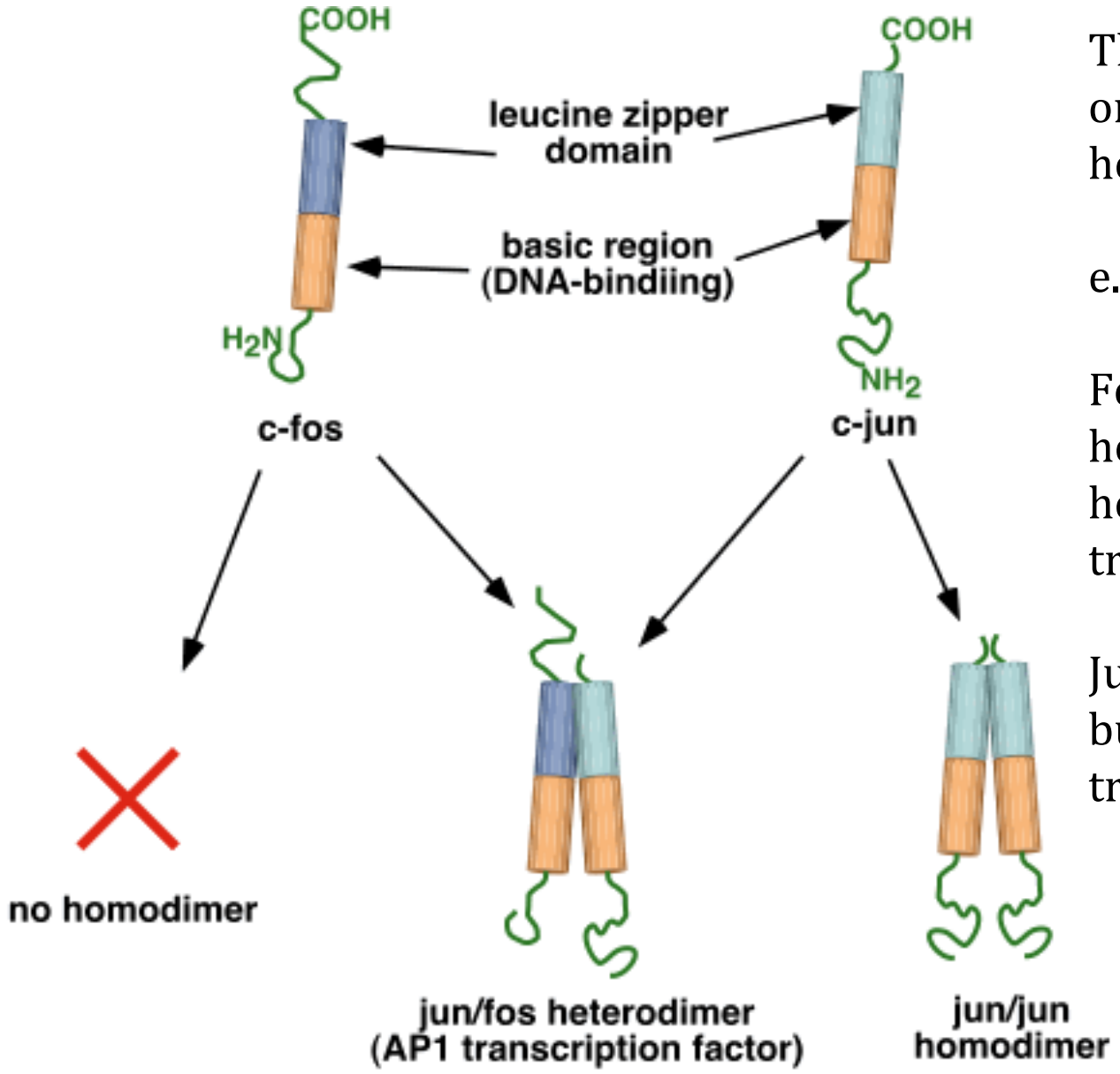








Classification of protein-DNA motifs



3. Leucine zippers

The Leu zippers are active only as dimers:
homo- or heterodimers

e.g. Fos and Jun

Fos cannot form a homodimer, but forms a heterodimer with Jun => transcription factor AP-1.

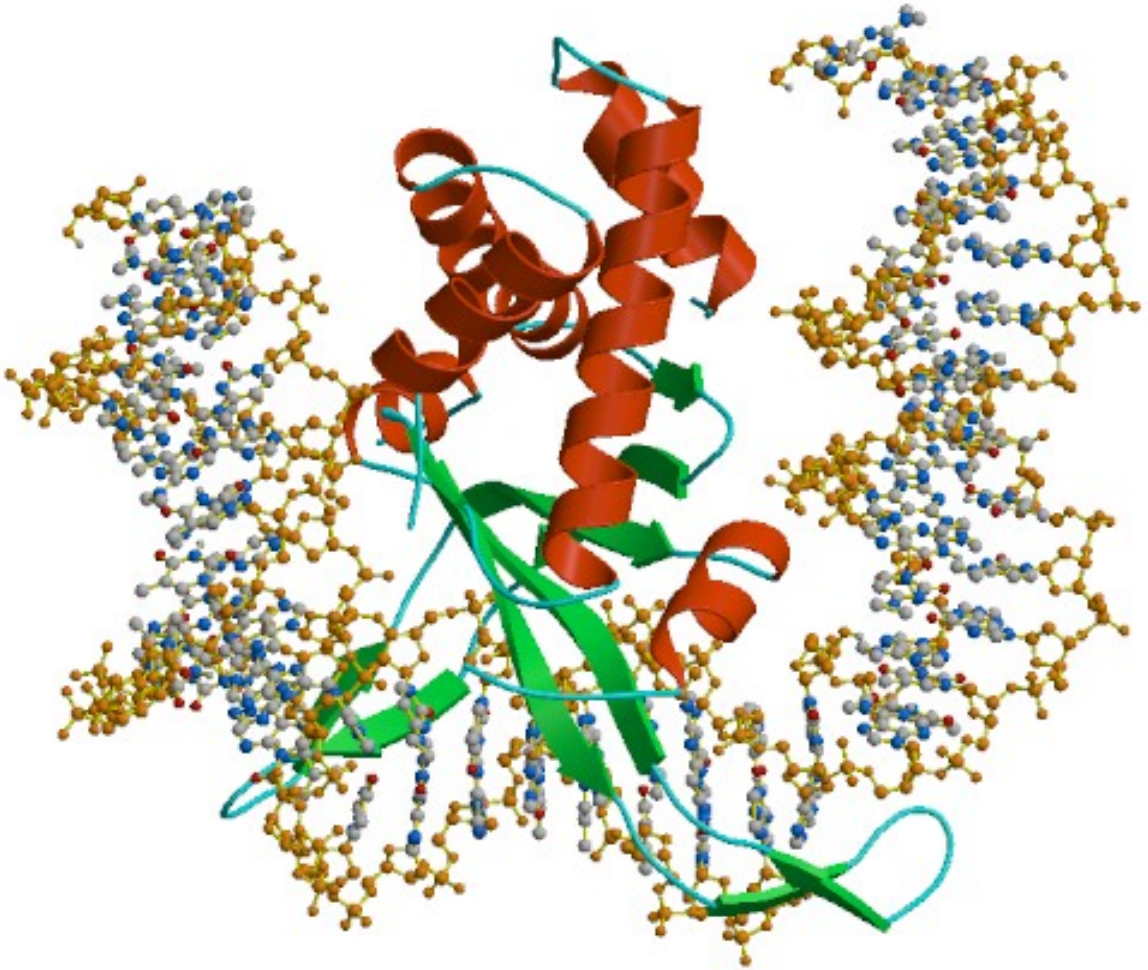
Jun can form homodimers, but they are not active in transcription.

The homo/ heterodimerization modifies the specificity of DNA binding

Classification of protein-DNA motifs

4. β-Ribbon group

Different recognition motif : binds to DNA via β-strands



Example: integration host factor
Folds DNA to compact it (cf histones)

The 'arms' of the protein enter into the minor groove, and expand it by partial intercalation of hydrophobic residues.

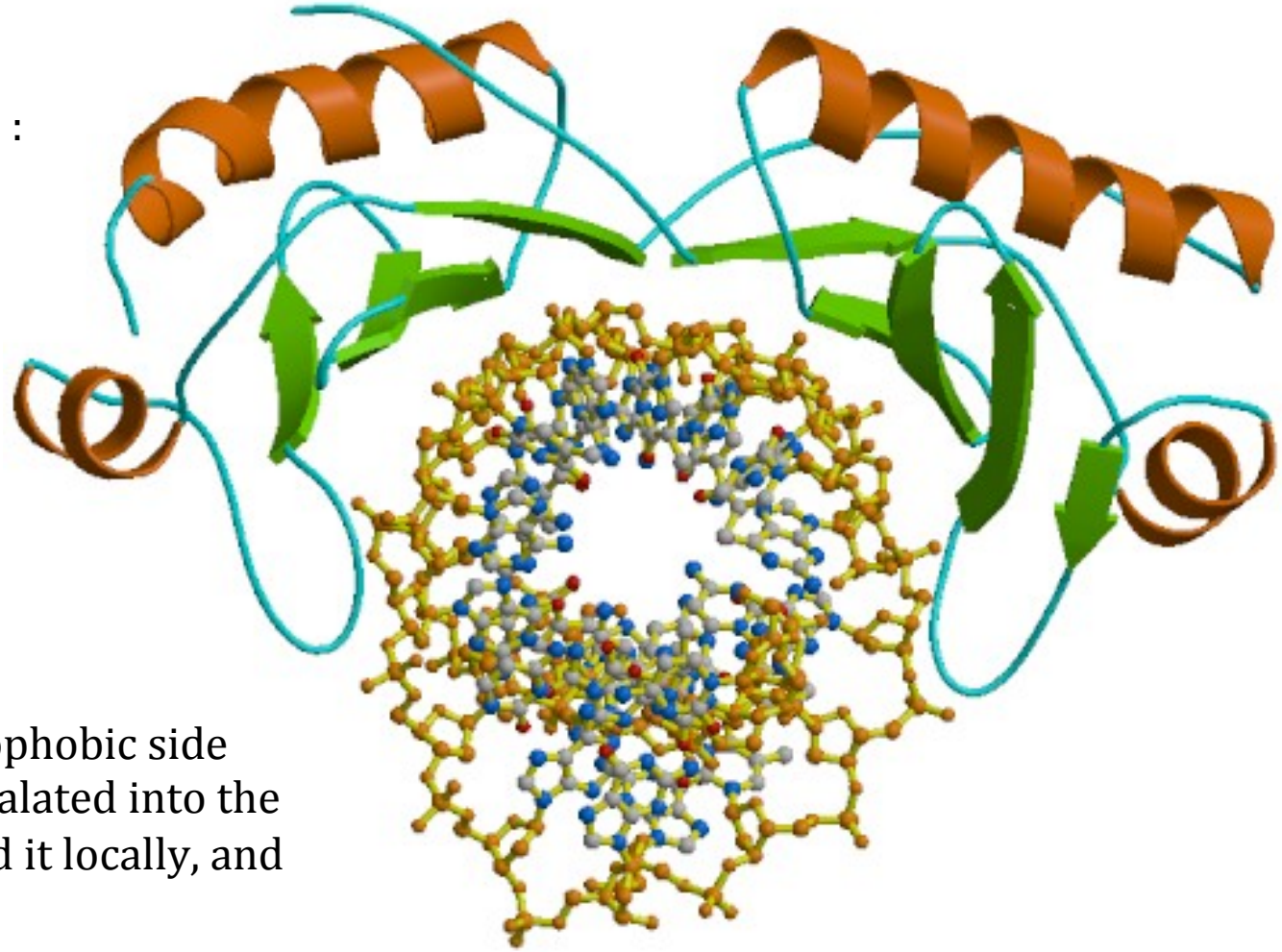
=> DNA wraps around the protein and β-strands are wound around the DNA.

Classification of protein-DNA motifs

4. β-Ribbon group

Other example : TATA box binding protein
Important role in transcription in eukaryotes

Recognition of DNA sequences rich in A and T : TATA(A ou T)A(A ou T)



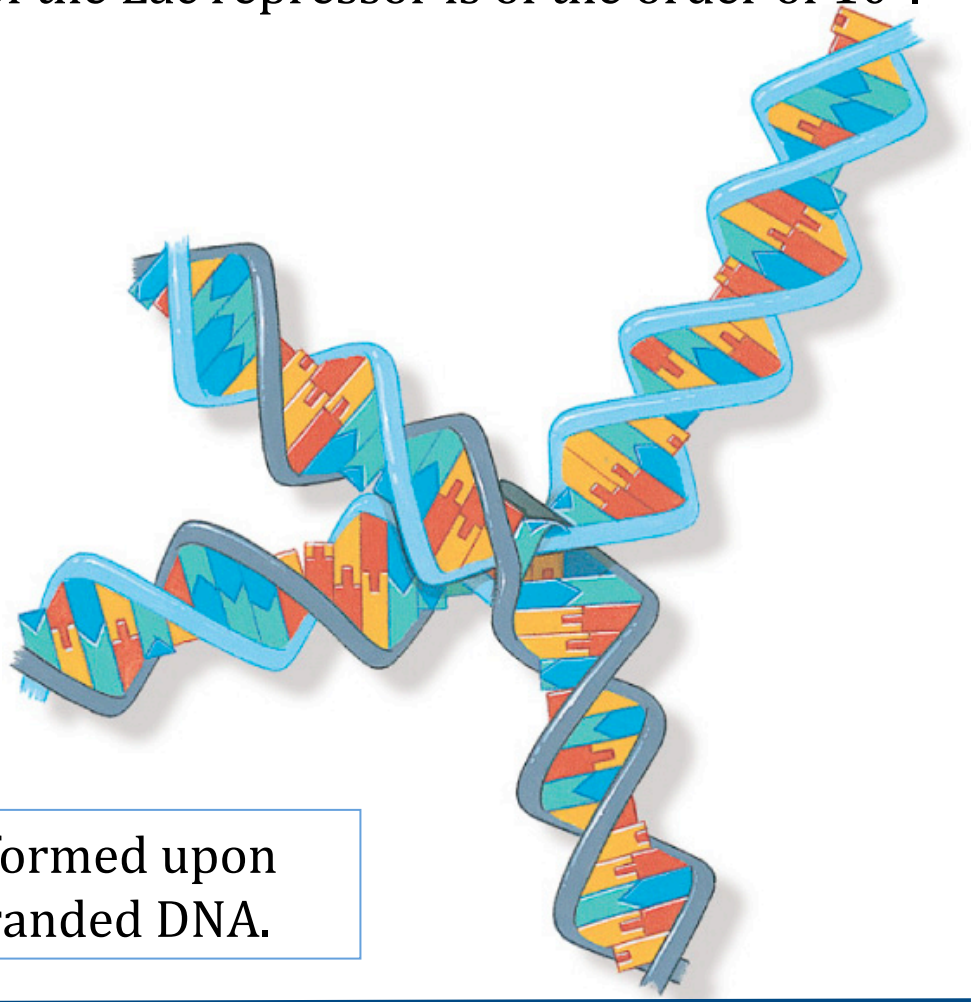
During recognition, hydrophobic side chains are partially intercalated into the minor groove. They uwind it locally, and bend it

Specificity of protein-DNA interactions

Many proteins which bind to DNA are able to recognize specific DNA sequences out of thousands of others
e.g. transcription factors: the specificity of the Lac repressor is of the order of 10^6.

At the other extreme, some proteins are not specific for the sequence - but specific to a certain structure ; e.g.

- repair enzymes recognize damaged DNA or DNA mismatches.
- Topoisomerases detect circular DNA
- The HMG1/2 bind to Holliday junctions.
- Histones are totally nonspecific, they compact DNA.



A Holliday junction is an intermediate formed upon recombination between two double-stranded DNA.

Specificity of protein-DNA interactions

Several mechanisms can make a DNA-binding protein specific for a given DNA target.

Level 1: intrinsic propensity of a motif to recognize a DNA sequence (via specific interactions)

Sometimes, proteins have additional contacts in the adjacent groove -> increases the affinity and specificity

e.g. in the HTH family:

- Homeodomains have an N-terminal 'arm' that is unstructured in the absence of DNA and form contacts in the minor groove adjacent to the major groove recognized by the recognition helix.
- In other HTH domains, like in the ets domains, small domains of β-strands bind in the minor groove (called winged HTH).

=> Increase of the target DNA sequence => increased specificity

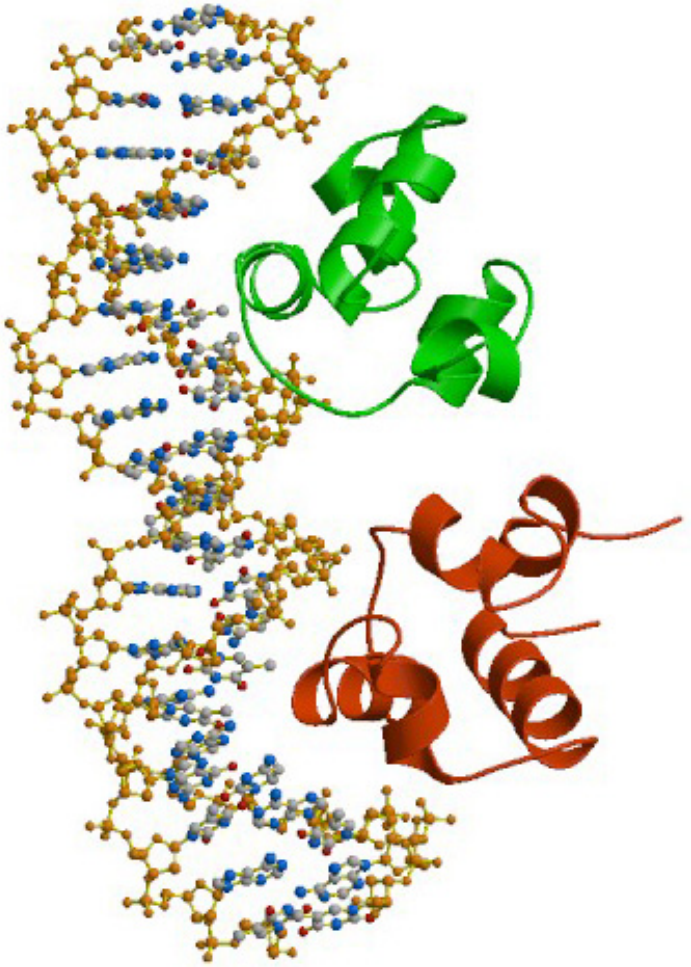
Level 2: specificity can be increased by assembling various DNA binding motifs.

Level 3: specificity can also be acquired indirectly, when non-specific proteins bind to specific proteins

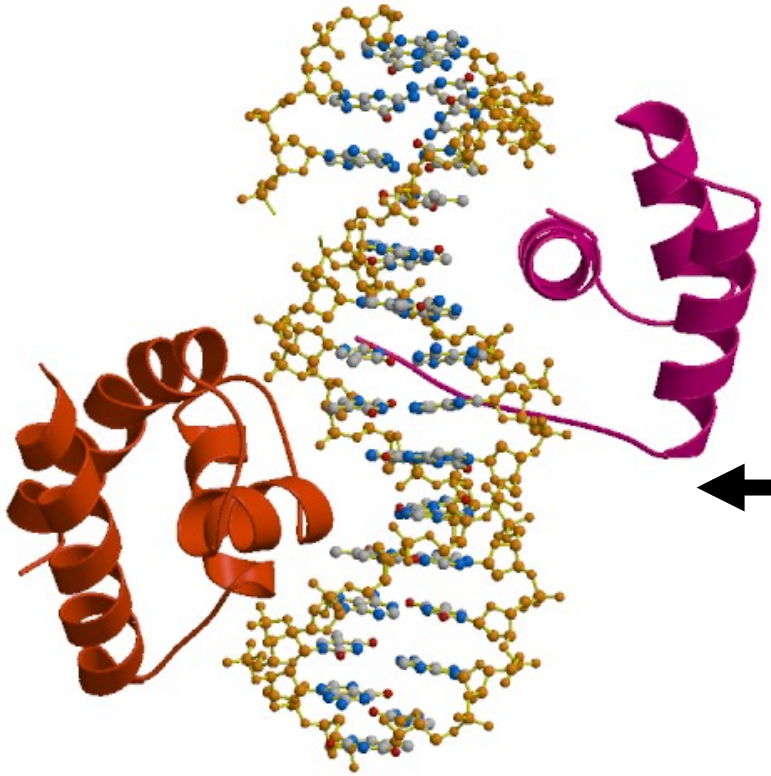
** The specificity is also related to the propensity of local deformation of DNA (A-T are sequences more flexible)

Specificity of protein-DNA interactions

Example: 2 identical domains => recognition site of DNA has two half sites that are either identical or palindromic.



The base pair separation between the 2 sites is specific for each protein: depends on the protein dimer interface



Combination of two different HTH motifs (here oci1 and pou) => association of two DNA recognition sites; includes 2 major grooves and 1 minor groove.

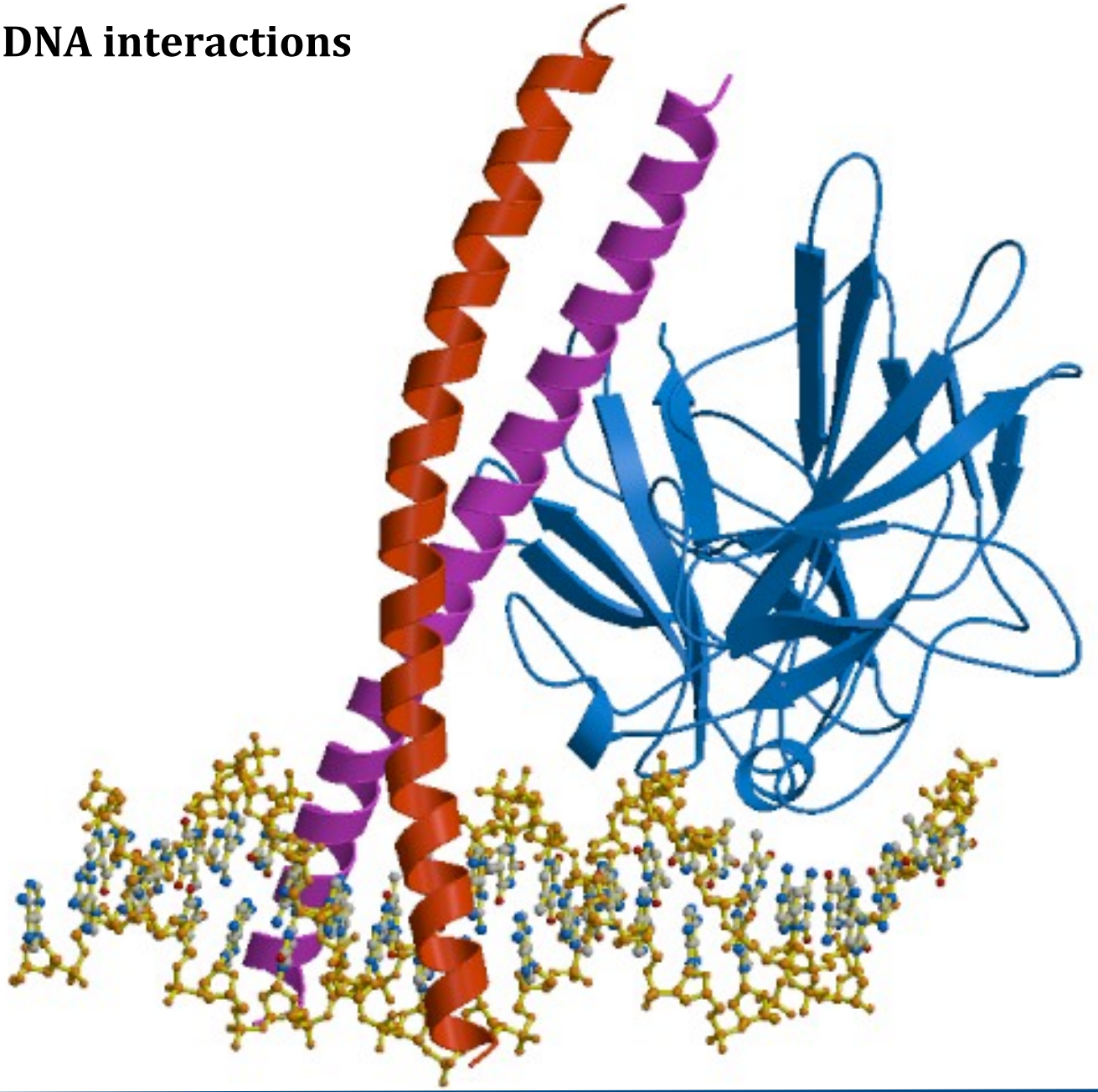
Specificity of protein-DNA interactions

Association between two different DNA-binding domains :

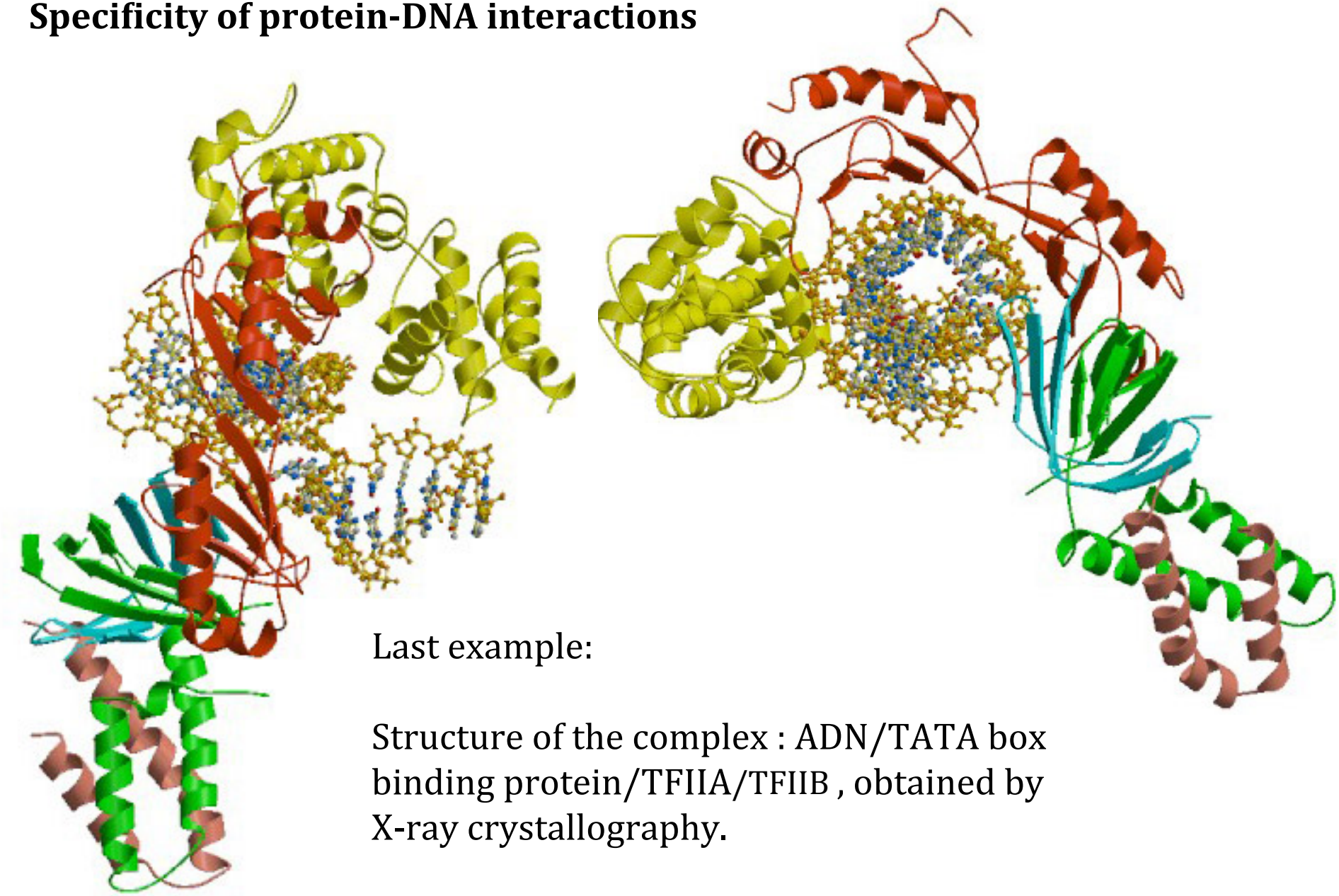
- Leu zipper (Fos and Jun ; purple and red)

- NFAT (blue)

The recognition sequence comprises 3 major grooves and 1 minor groove.



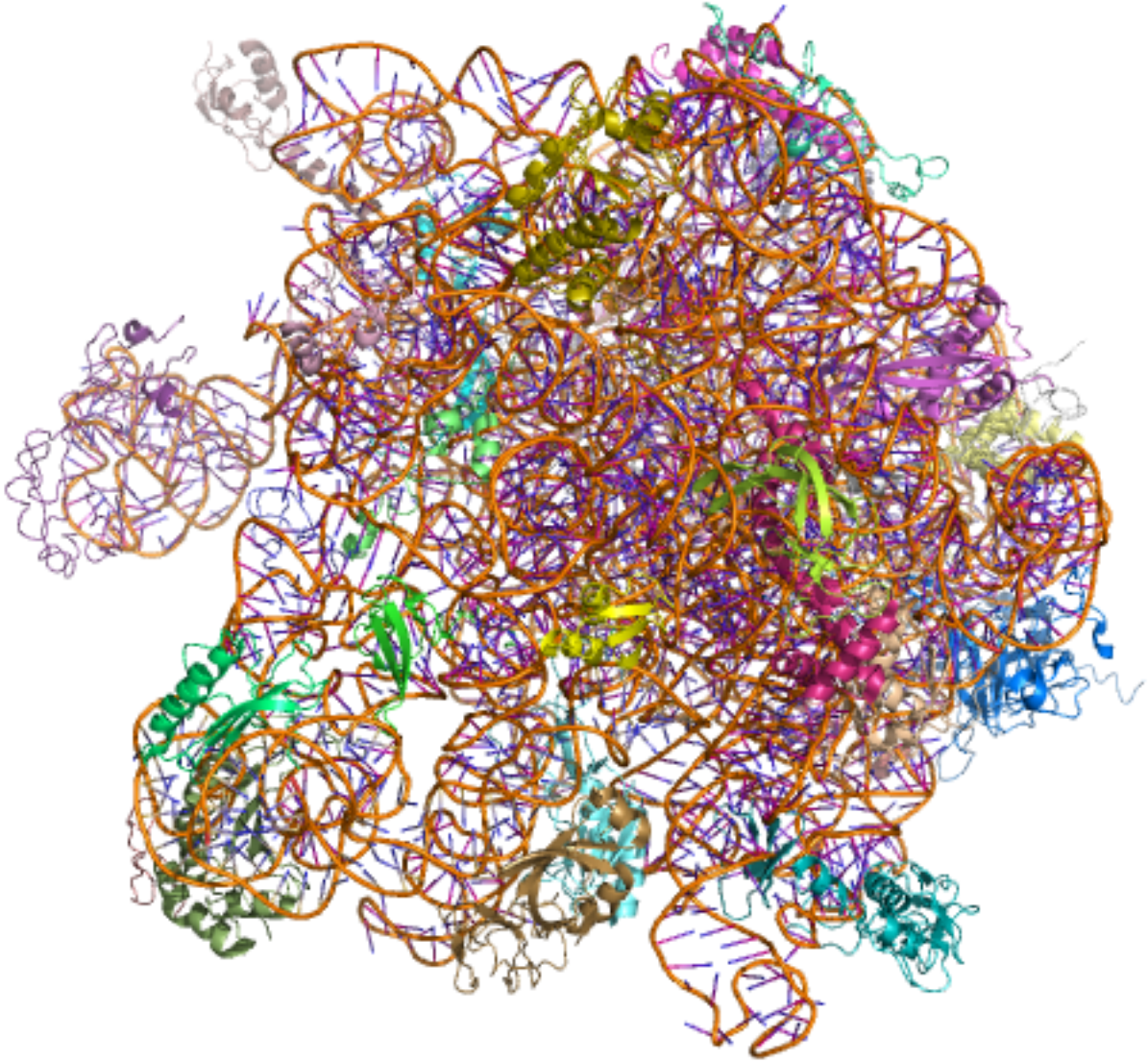
Specificity of protein-DNA interactions



Last example:

Structure of the complex : ADN/TATA box binding protein/TFIIA/TFIIB , obtained by X-ray crystallography.

Protein-RNA interactions



Ribosome: translation mRNA -> protein

Complex with thousands of nucleobases and dozens of proteins

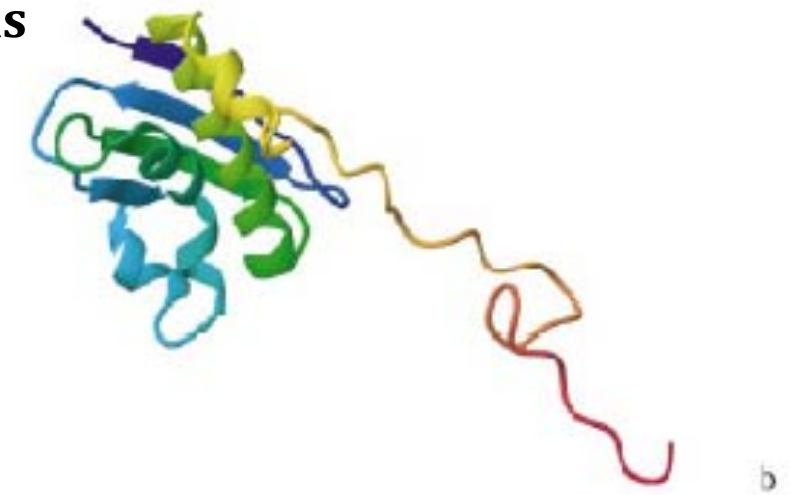
Basic principles identical for protein-DNA and protein-RNA.

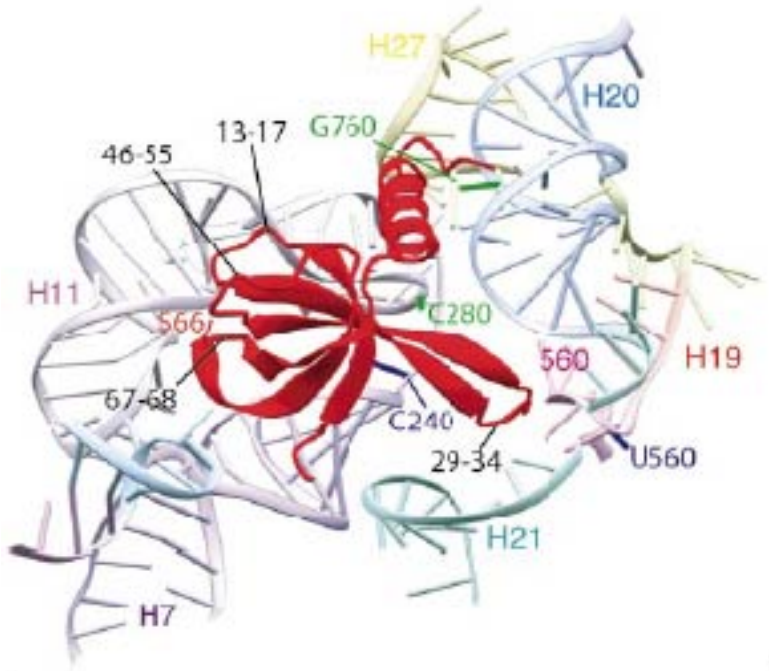
But differ because of DNA-RNA differences:

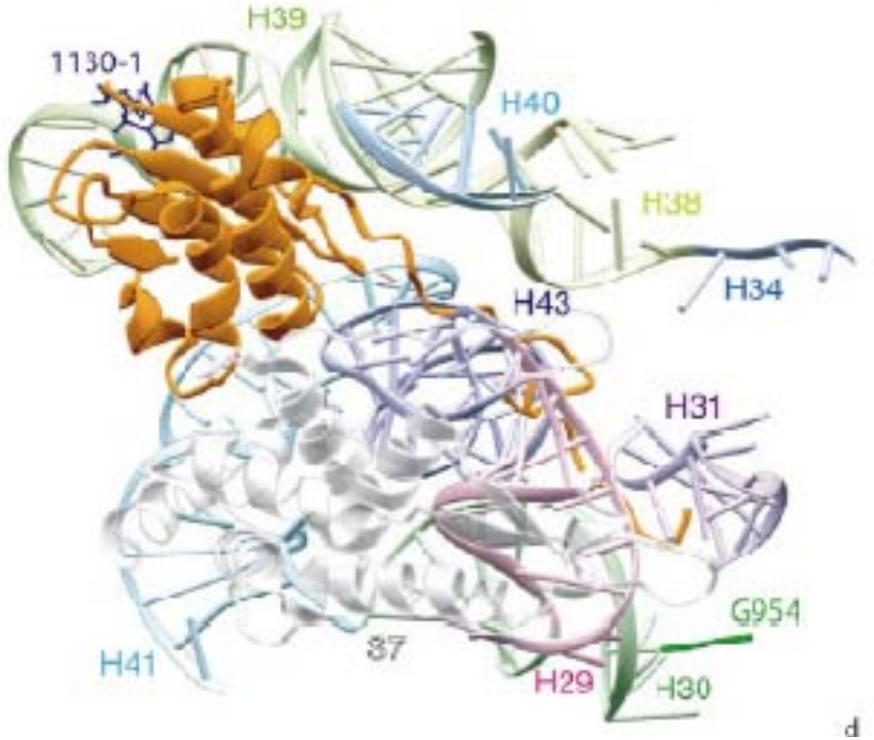
Flexibility …

Protein-RNA interactions

Larger view of protein-RNA interactions



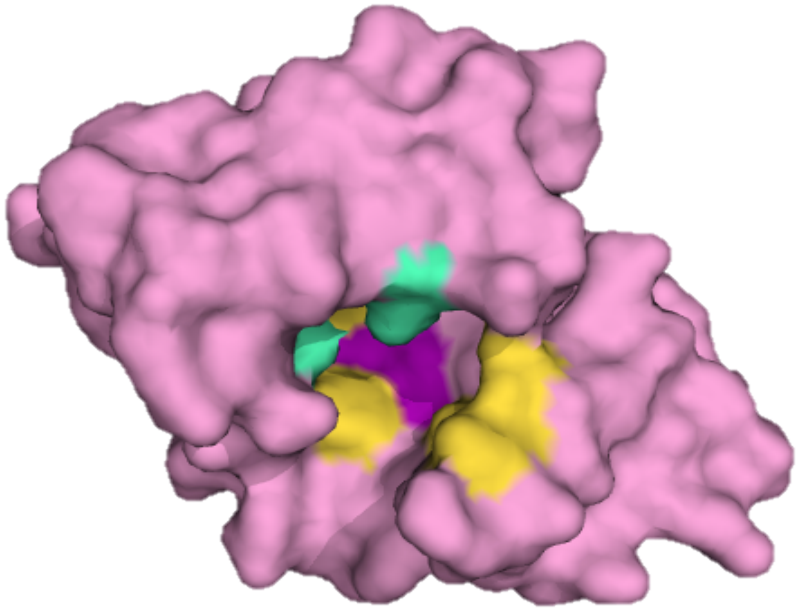


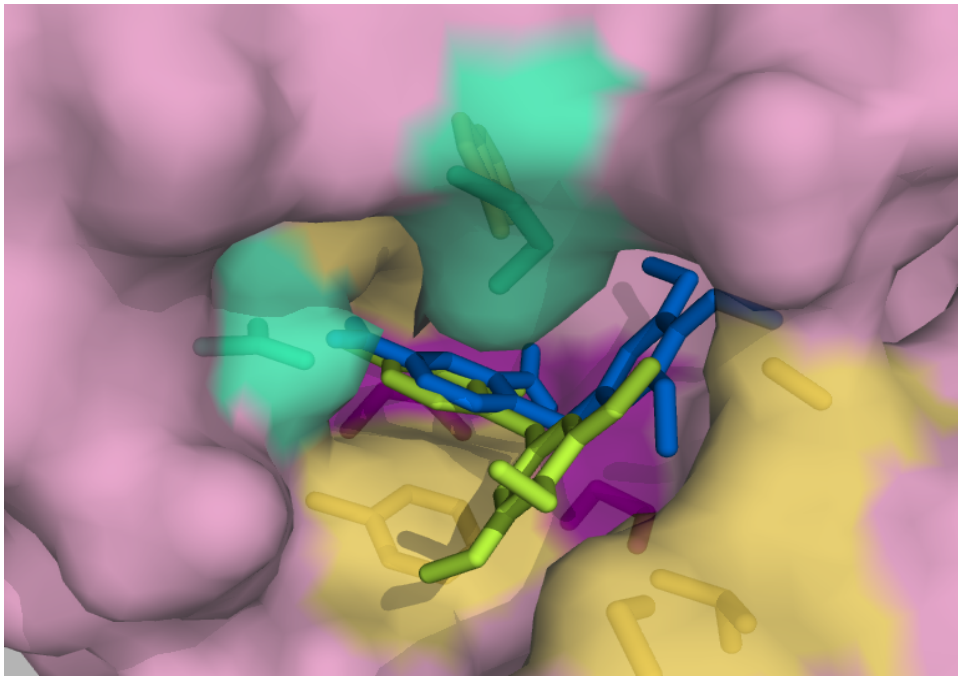


Protein-ligand interactions

Ligand : peptide or small organic molecule

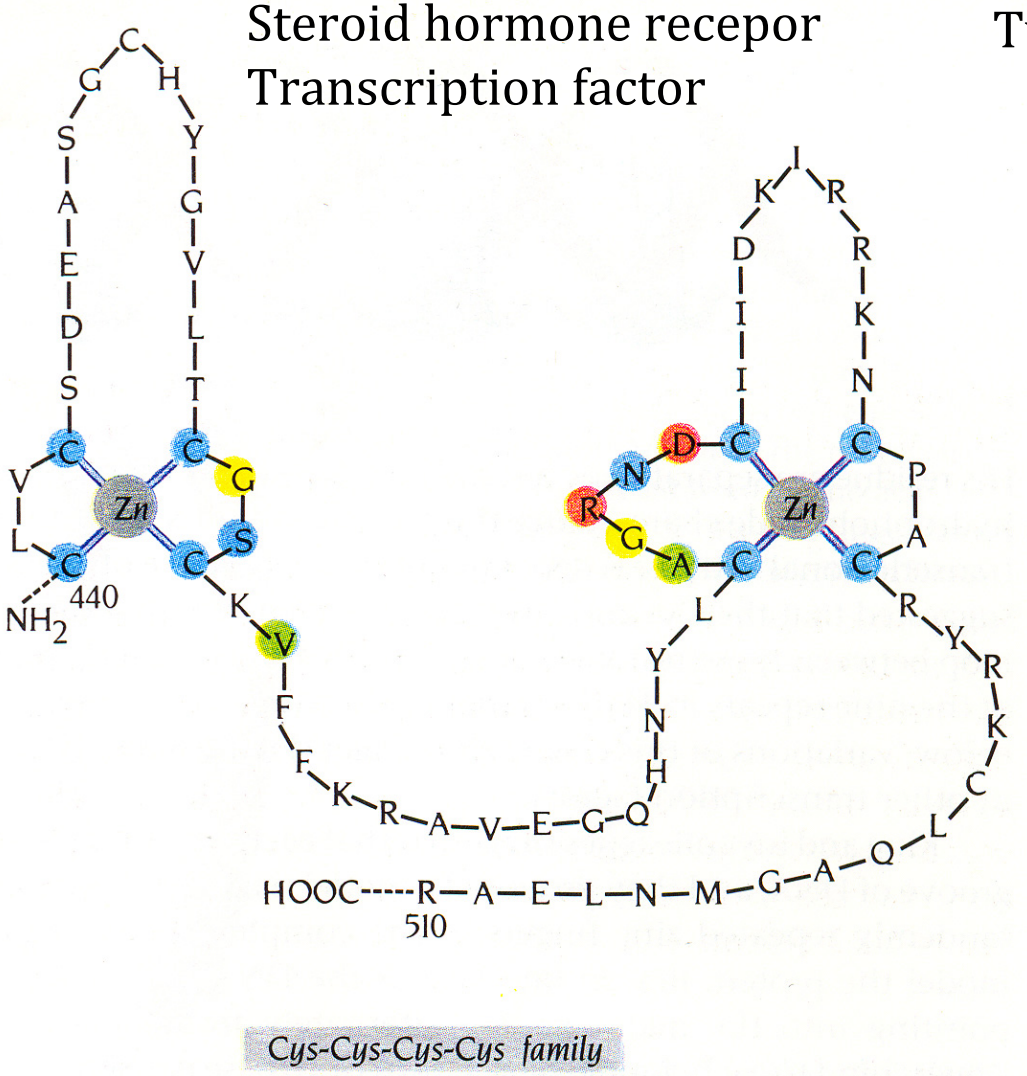
Allows the activation or inhibition of the receptor protein
→Modulation of the activity (enzymatic, or DNA-binding, …)
→ Can be used as drug
→ -> -> Rational drug design





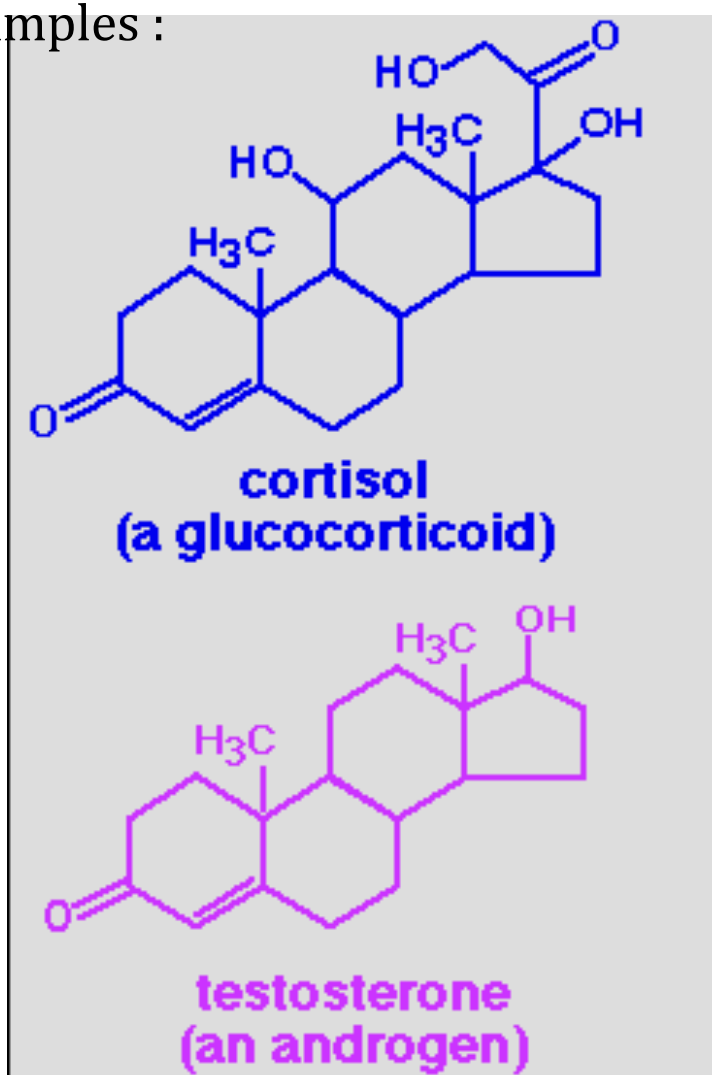
Protein-ligand interactions

Example : Zinc finger:
Steroid hormone recepor
Transcription factor



Cys-Cys-Cys-Cys family

Steroid hormones (4 cycles of 17 C atoms) – important in higher eukaryotes – hydrophobic molecules – Two examples :



Protein-ligand interactions

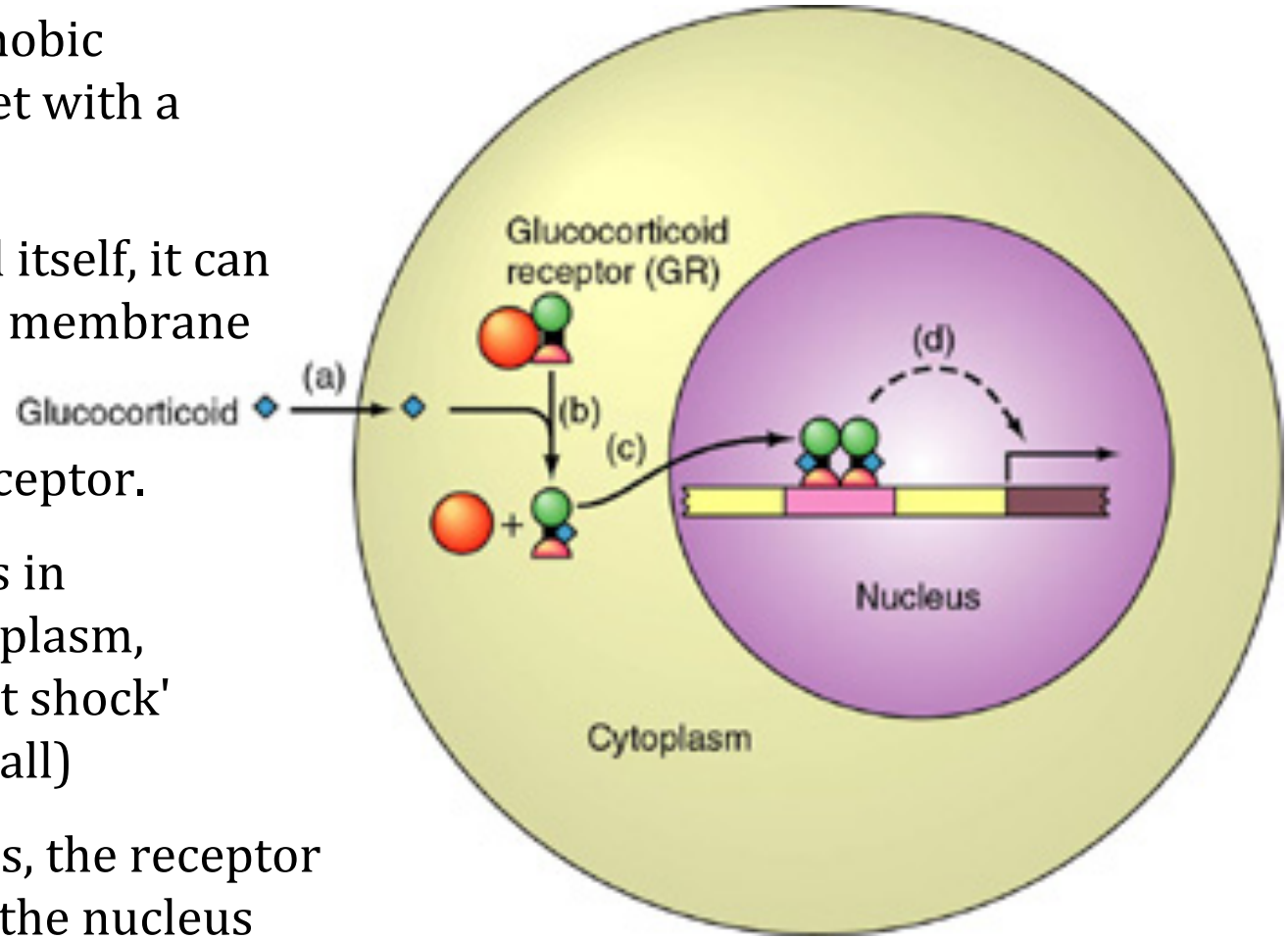
The hormone is hydrophobic
=> it must reach its target with a transport molecule

On the surface of the cell itself, it can diffuse through the lipid membrane

In the cytoplasm, it must be handled by a receptor.

The receptor (protein) is in inactive form in the cytoplasm, in complex with the 'heat shock' hsp90 protein (orange ball)

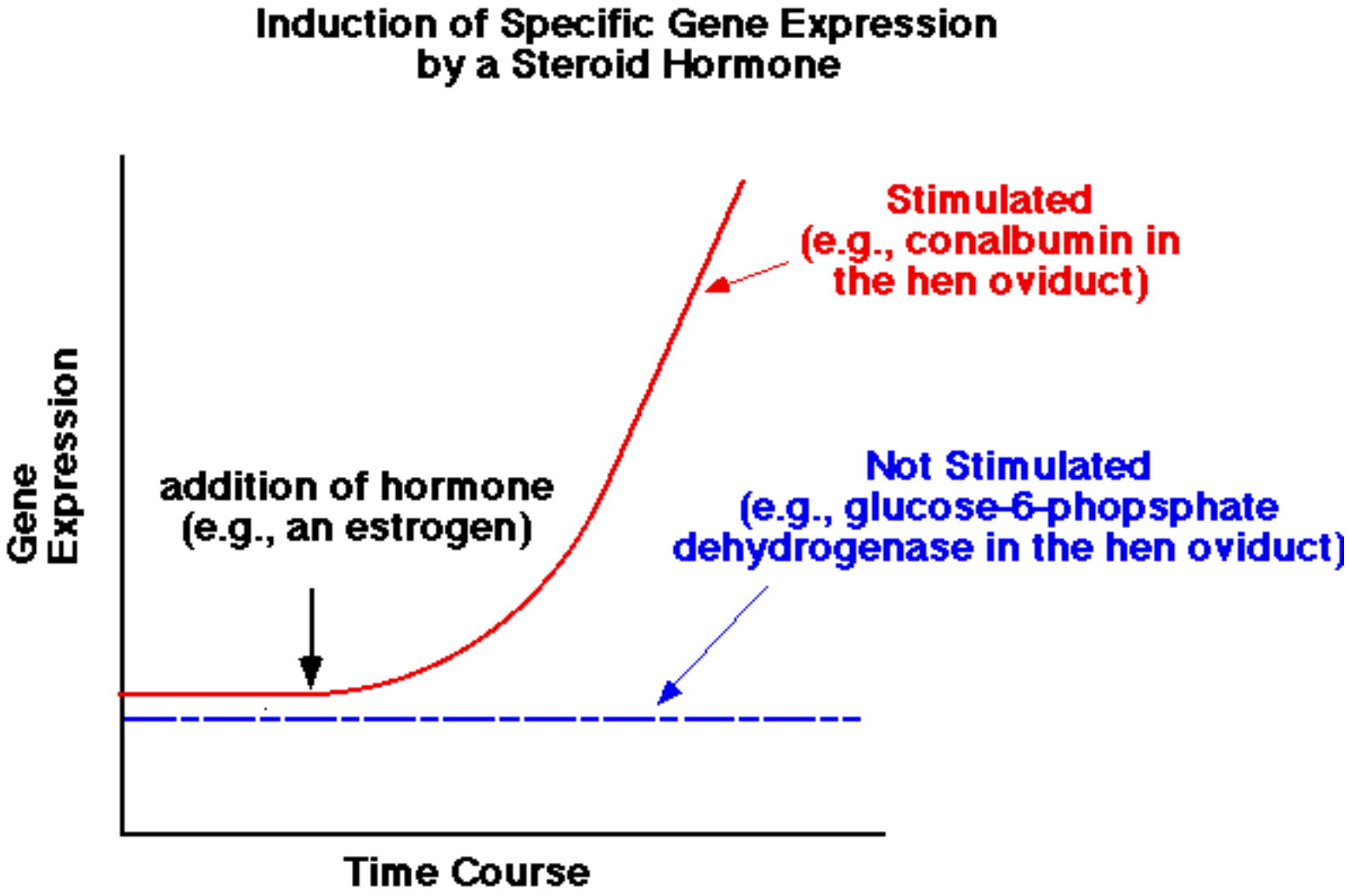
When the hormone binds, the receptor is activated, and goes to the nucleus



There, the Zn finger domain binds to specific DNA sites, and gene expression is activated.

Protein-ligand interactions

Response of hormone-sensitive a tissue to treatment with steroid hormones (estrogen here) - here in hen oviduct

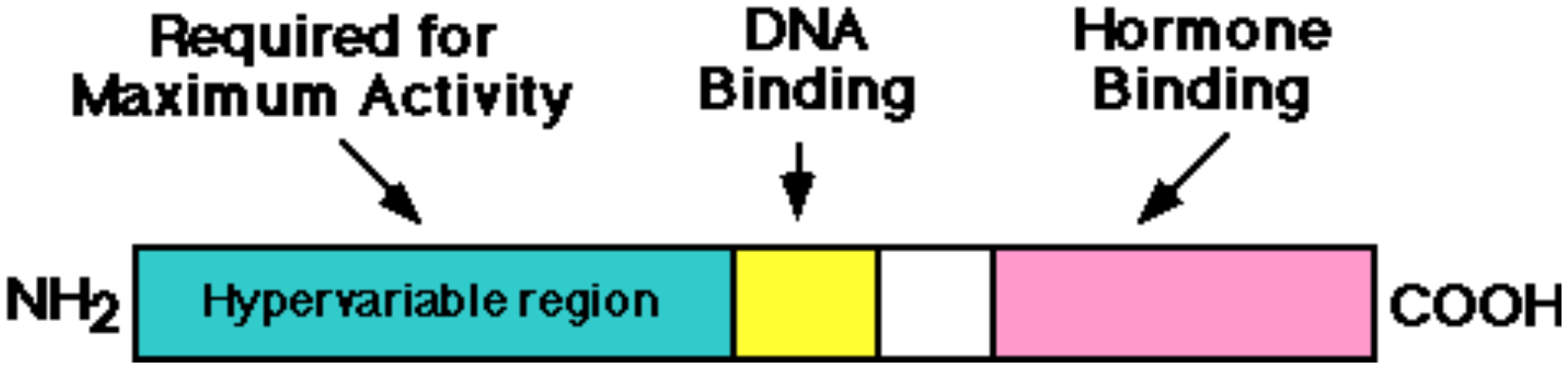


Protein-ligand interactions

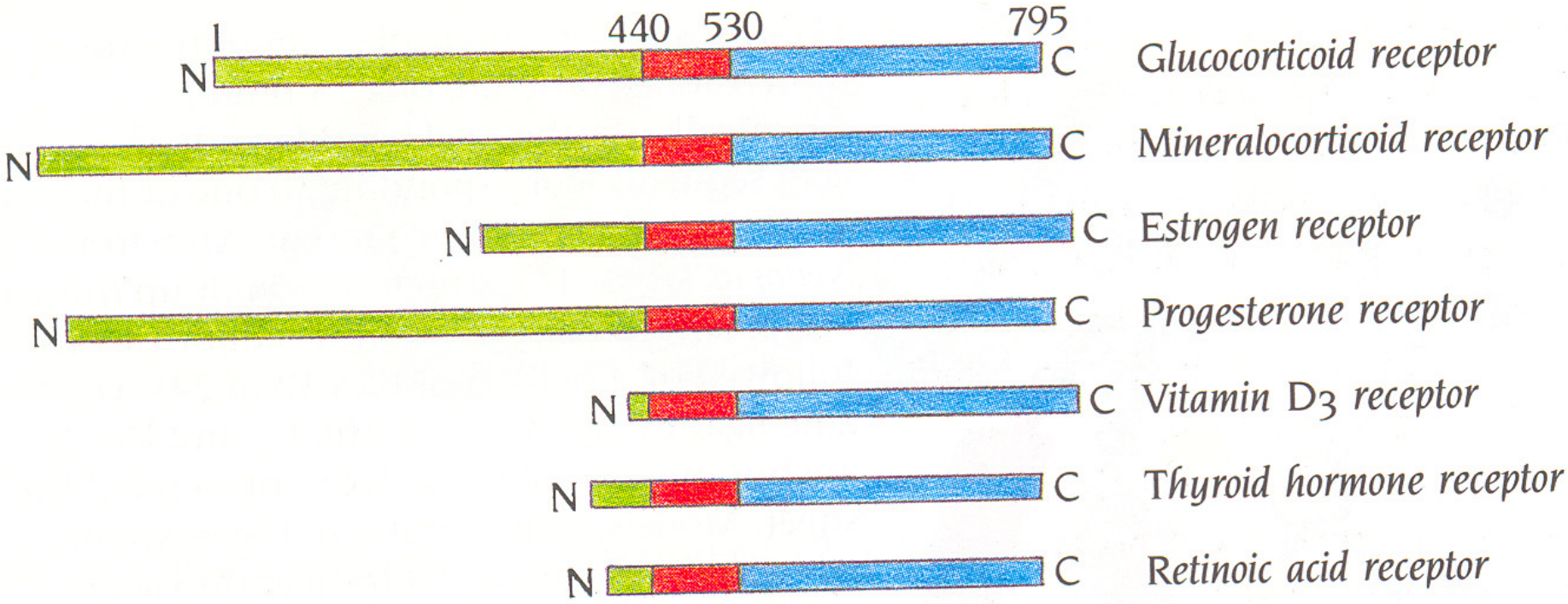
The hormone receptors are a class of proteins activated by a ligand. When bound to DNA, they serve as on/off switch for the transcription of certain genes in the nucleus.

-> Control of development and differentiation of the skin, bones and behavioral centers in the brain, and continuous regulation of tissues related to reproduction.

Several domains with different roles:
- Variable N-terminal region,
- Conserved DNA-binding domain,
- Variable hinge domain
- Conserved ligand binding domain
- Variable C-terminal region.



Protein-ligand interactions



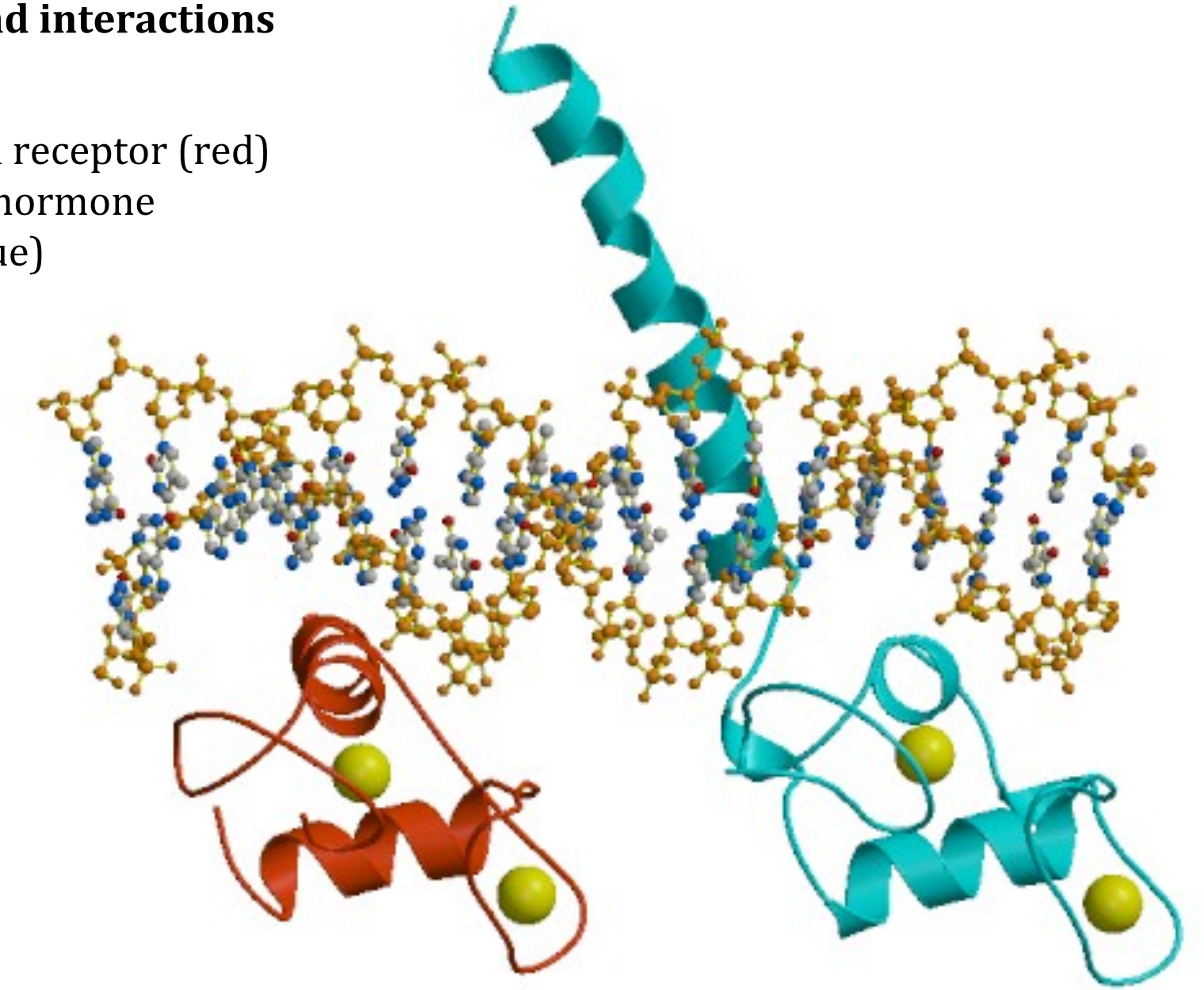
Family of receptors, linked by evolution (Zn fingers)

The DNA-binding domains (red) have very similar amino acid sequences, whereas the hormone-binding domain (blue) is more variable.

Suggests that different domains can act as independent domains – and can be interchanged between family members

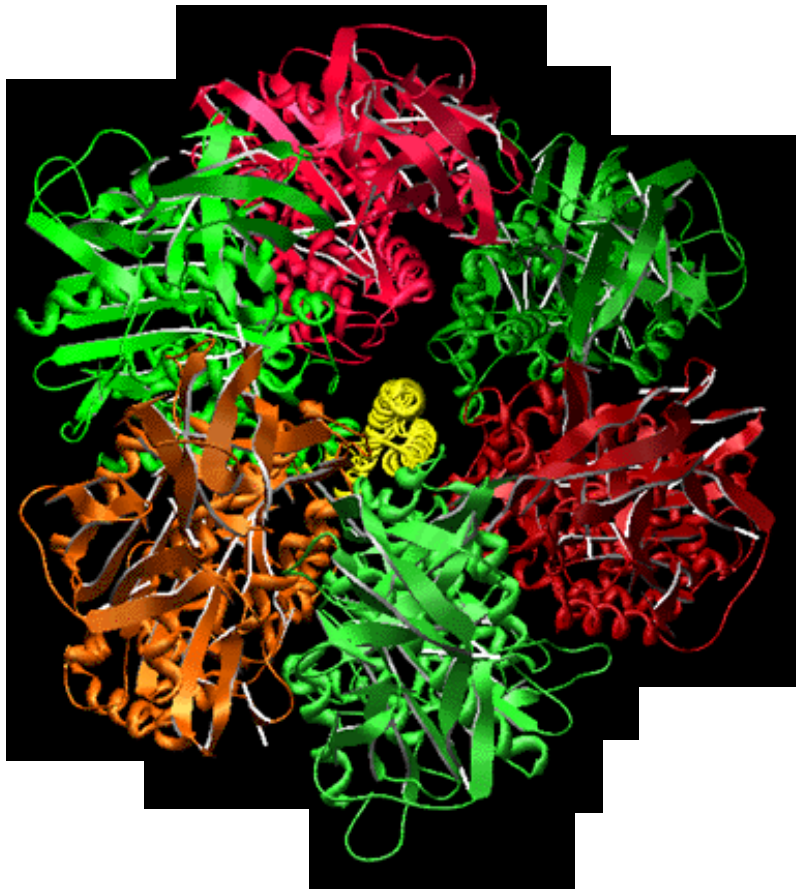
Protein-ligand interactions

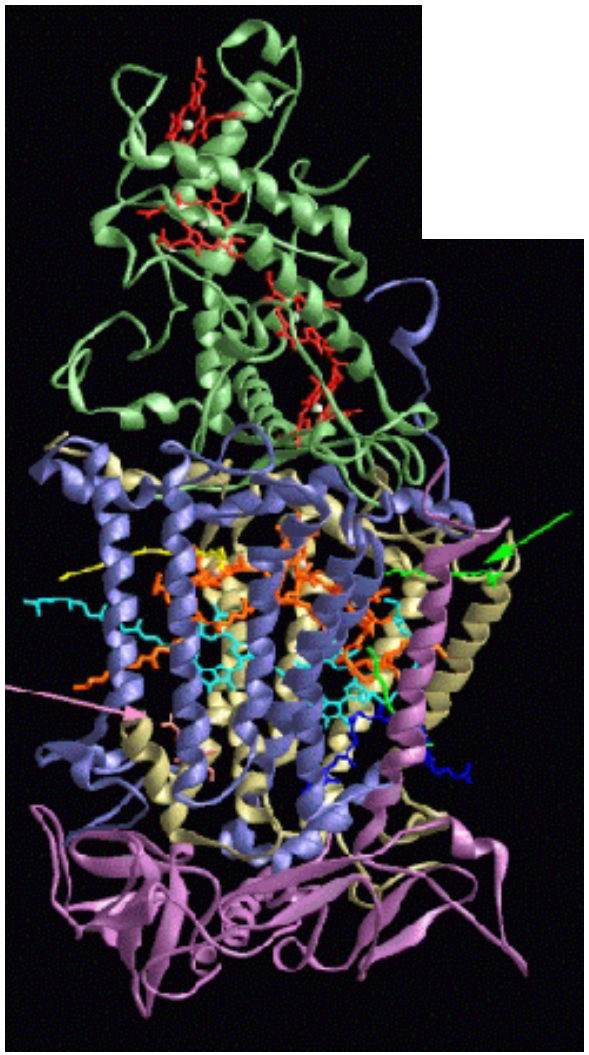
Retinoic acid receptor (red) and thyroid hormone receptor (blue)



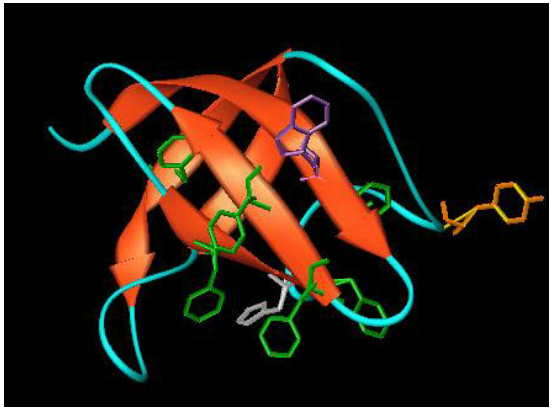
Protein structures -> classification

Secondary structure -> tertiary structure -> quaternary structure -> function

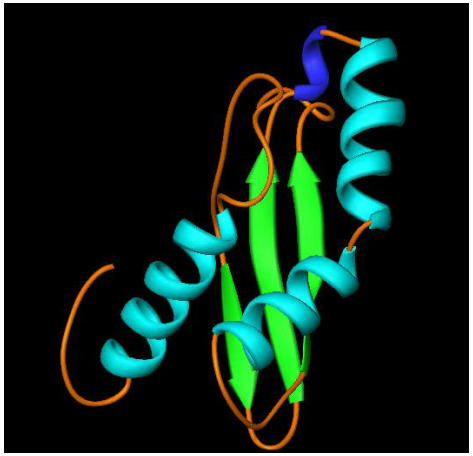




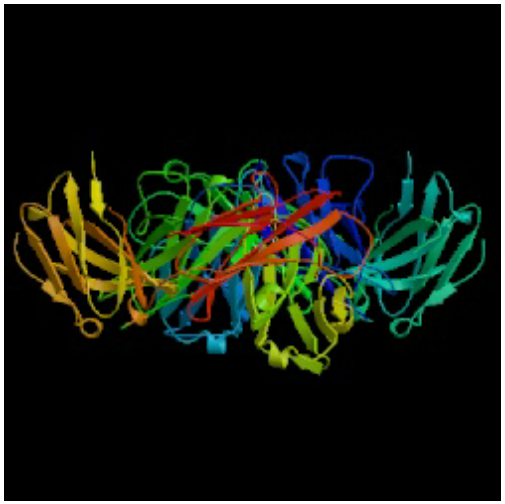
Examples of protein structures



Cold shock protein

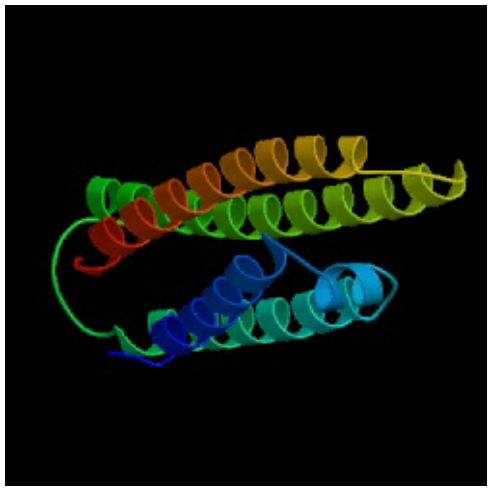


Ribosomal binding factor



Immunoglobulin

Triose Phosphate Isomerase dimer



Apolipoprotein E3

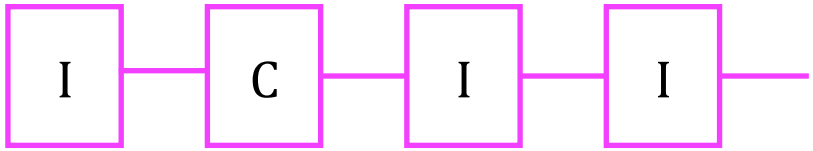
Myoglobin

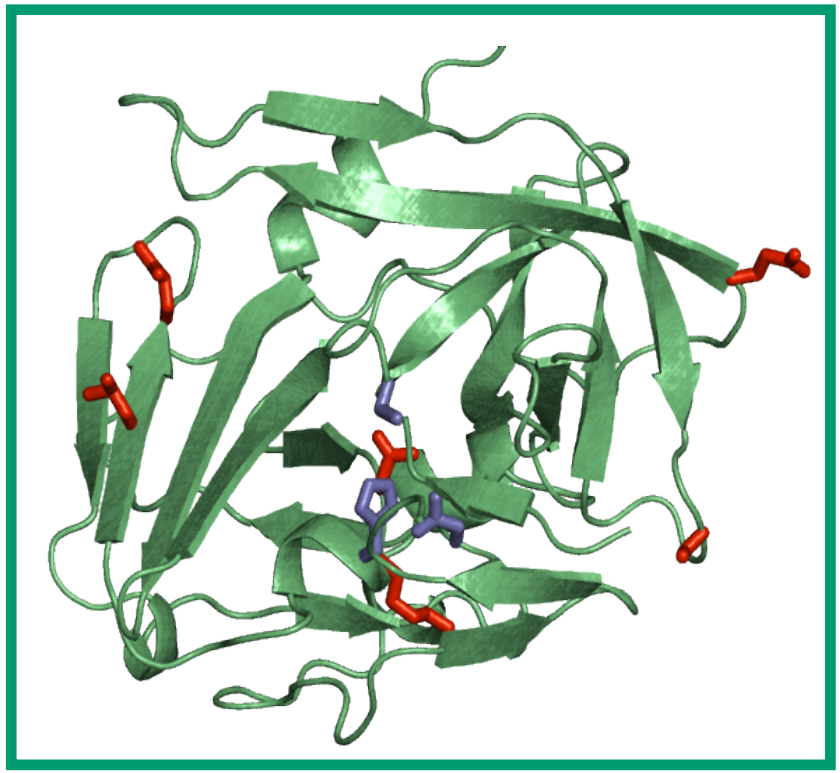
Protein structures

Sequence

3D structure

Biological function:
catalysis, regulation,
transport, ...

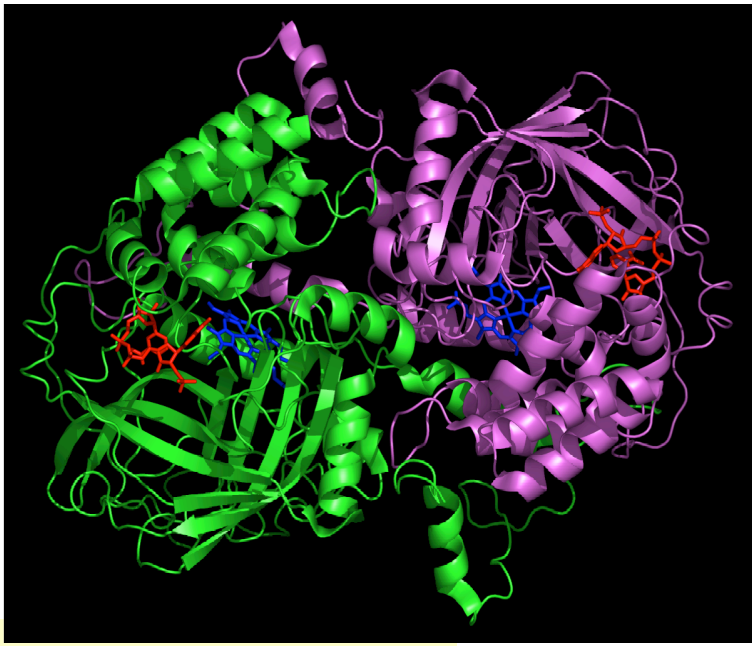


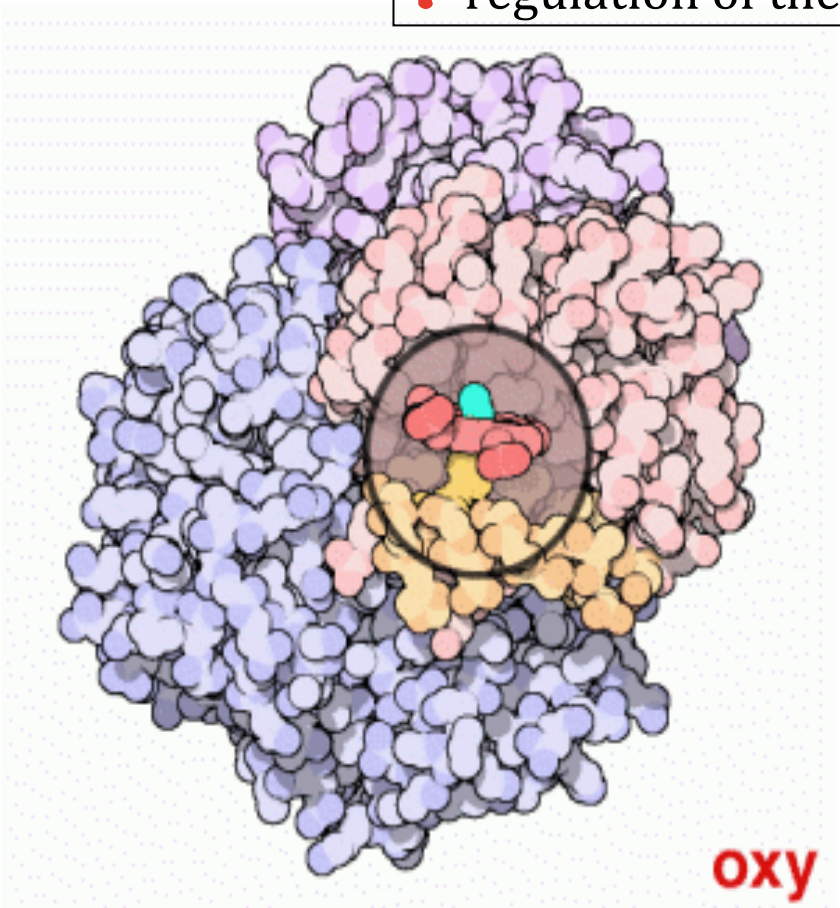


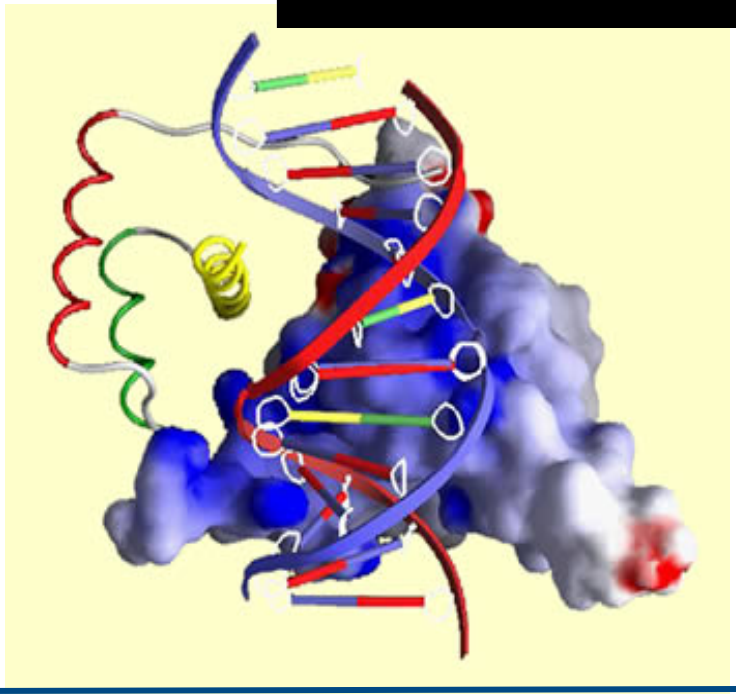
In general:
- one sequence -> one 3D structure (! exceptions!)
- this 3D structure is the absolute free energy minimum.

Protein structures

Proteins have various roles
- catalysis of biochemical reactions (enzymes)
- transport of molecules or ions (ex. hemoglobin)
- regulation of the genome







Protein structure comparisons

When protein structures are available, the alignment of structures is a better/ more reliable tool than sequence alignments for the detection of homology.

Because structures have the tendency to diverge less than sequences. Proteins with some sequence similarity have usually very similar structures. The determination of residue-residue correspondences by structure alignment is a powerful method of sequence alignment.

Generally above 40% sequence similarity, the structures are very similar. There are cases where the sequence similarity is low (20-25%), while the structural similarity is still high.

Aligning/superimposing structures is used to identify regions that have identical structures, to classify these structures, and group them into families.

More powerful approach to detect homologies that sequence alignment. But not always applicable as there are much more sequences than structures ...

Protein structure comparisons

Measure of structural similarity

The most commonly measure used is the rms or rmsd (root mean square deviation) after atomic superposition :

- - Consider a certain type of atoms (all atoms, main chain atoms , Cα, …)
  - Search for the corresponding atoms in the two proteins
  - Look for the superposition of structures that minimizes the rmsd = average distance between corresponding atoms after superposition of the structures:

$$\text{rms} = \underset{\substack{\text{all rotations} \\ \text{and translations}}}{Min} \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left(x_1^i - x_2^i\right)^2 + \left(y_1^i - y_2^i\right)^2 + \left(z_1^i - z_2^i\right)^2\right)}$$

Where $\left(x_1^i, y_1^i, z_1^i\right)$ and $\left(x_2^i, y_2^i, z_2^i\right)$ are the Cartesian coordinates of the corresponding atoms i in proteins 1 et 2 ;
N is the total number of atoms considered.

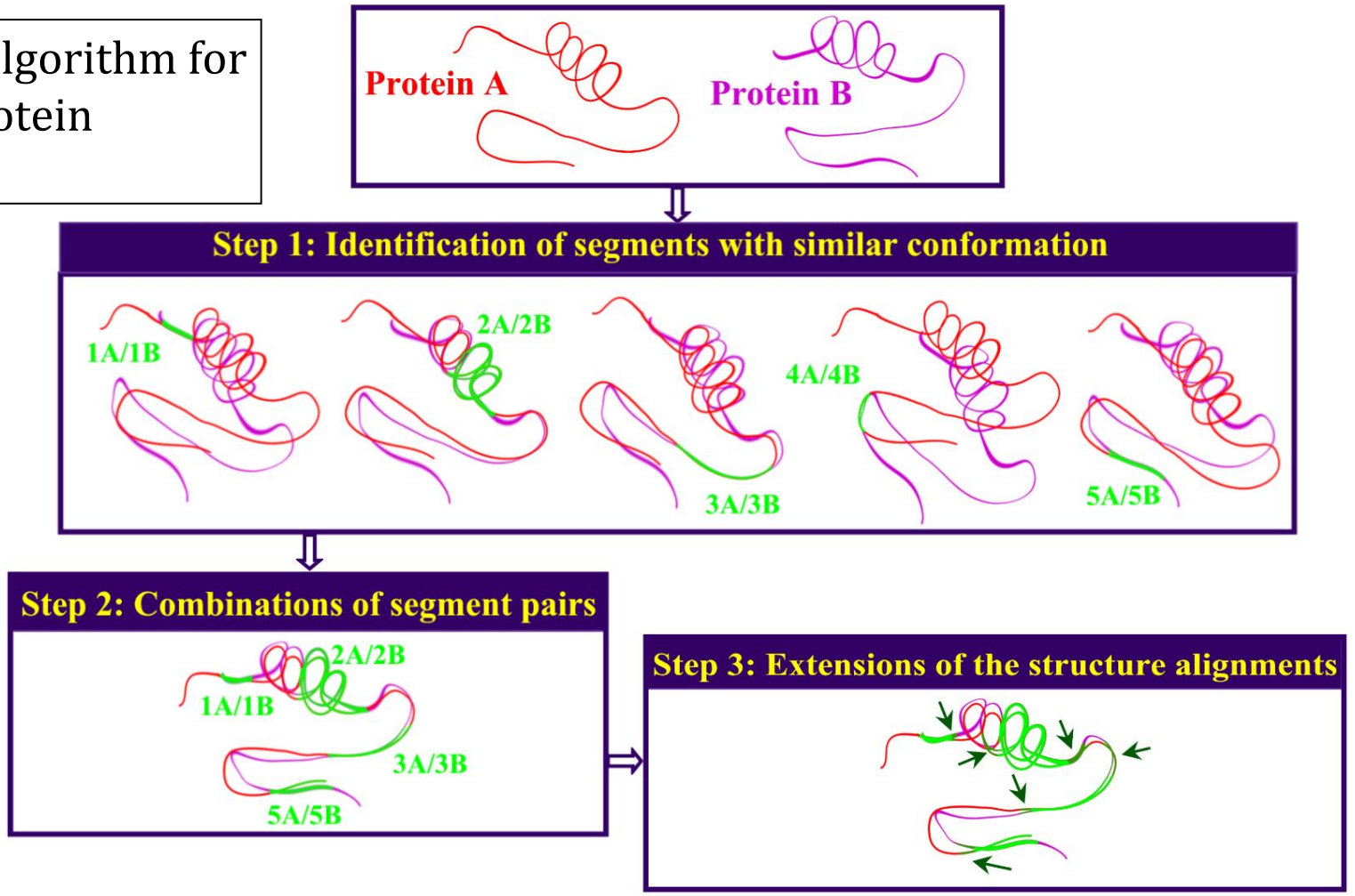
** Superpositions: translations et rotations of rigid bodies **

Note: the result vary according to the correspondence between the atoms, and to the atoms considered.

Protein structure comparisons

But how to find the correspondence between atoms, residues??
Using sequence alignments? NO ! More informative to use automatic structure comparisons.

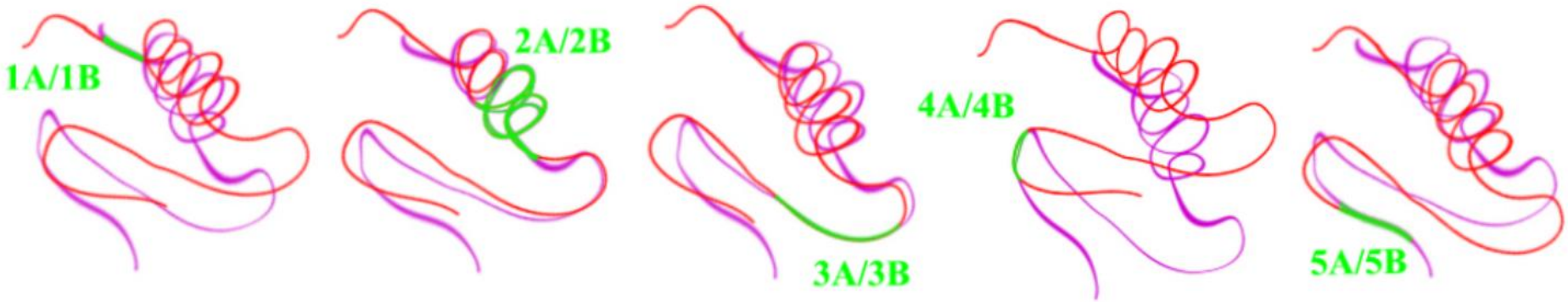
Example of an algorithm for aligning two protein structures:



Protein structure comparisons

Step 1: Identification of protein segments of similar conformations:

Similar local structure and/or secondary structure



To do this, each protein is divided into:

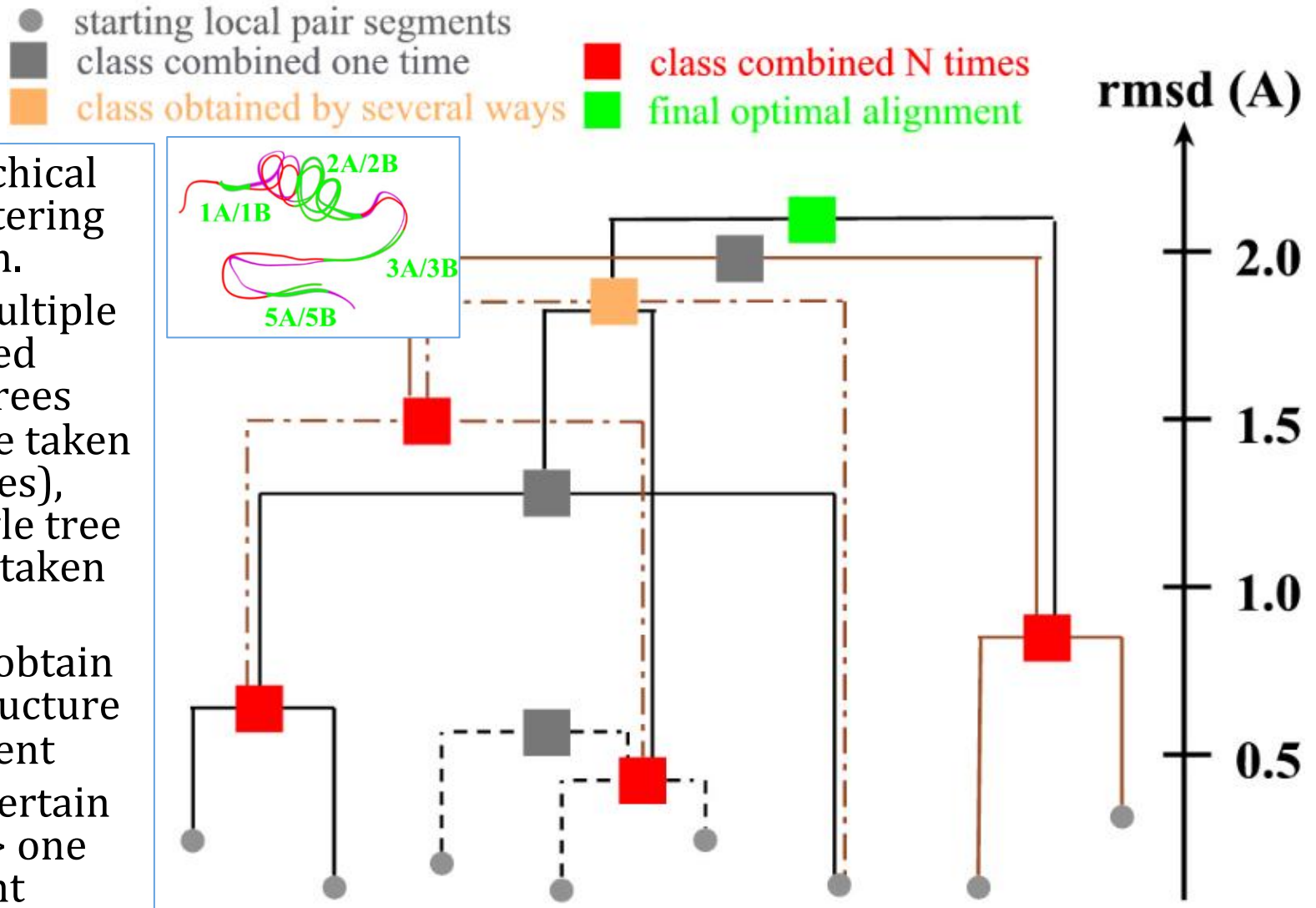
- Overlapping segments of n residues (n=5, .., 10,..)
- or Elements of secondary structures
- or Portions of secondary structures elements

The similarity of the segments is estimated by:

- Filter on the end-to-end distance (to increase calculation speed)
- Filter on the distance from the N-terminal end (to increase speed)
- Threshold value of the rms

Protein structure comparisons

Step 2: Combination of segment pairs => Global structure similarity



Here: Hierarchical tree-like clustering algorithm.

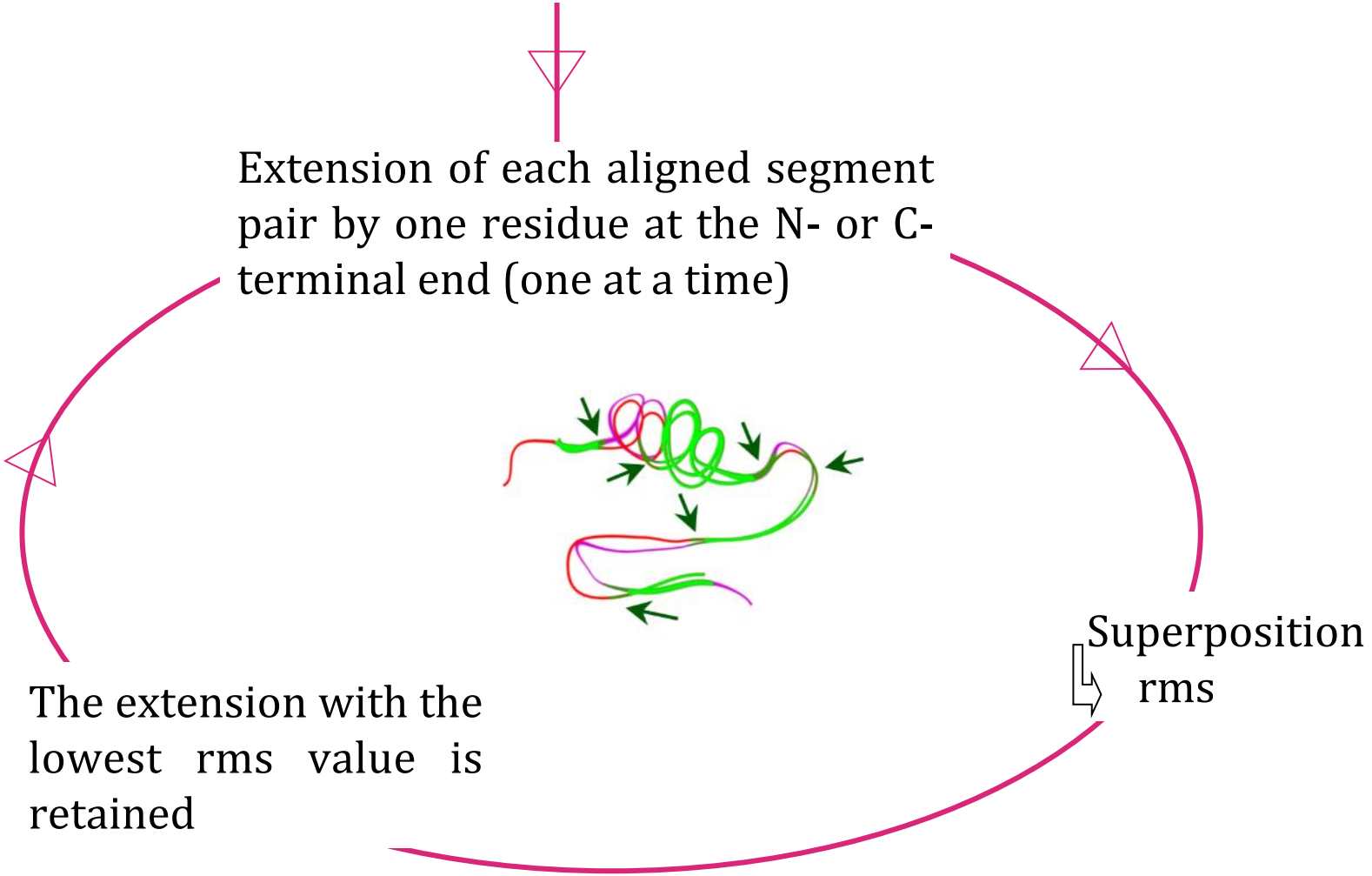
Peculiarity: Multiple intertwined clustering trees (classes may be taken several times), instead of single tree (each class is taken once)

⇒ needed to obtain optimal structure alignment

Cut tree at a certain rms value => one alignment

Protein structure comparisons

Step 3: Extension of the aligned regions : improving the alignment – when the segments are secondary structure elements: allows to include loops



Protein structure comparisons

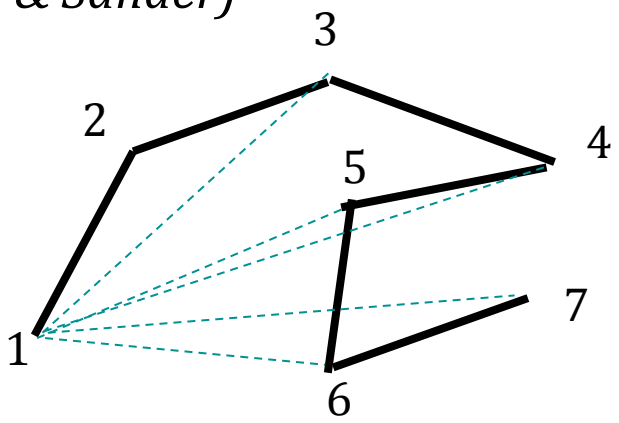
Example: Alignment of ubiquitin and ferredoxin

rms = 2.64 Å, 51 aligned residues

```
   EEEEEE      EEEEE       HHHHHHHHHHHH              EEEEE                  EE                EEEEEE
  MQIFVKTL TGKTITLEVEPSDTIENVKAKIQDKEGIPPD        QQRLIFAG                KQLEDGRTLSDYNIQKESTLHLVLRLRGG
  |||||||| ||||||||||||   ||||||||||           |||||||                 |||||||          ||||||
ATYKVTLINEAEGINETIDCDDDT   YILDAAEEAGLDLPYSCRAGACSTCAGTITSGTIDQSDQSFLDDDQIEAGYVLTCV     AYPTSDCTIKTHQEEGLY
   EEEEE      EEEEE         HHHHHHH                 EEEEEEEEE        HHHHH  EE         EEEEEEEEE
```

Protein structure comparisons

Other alignment algorithm: DALI (Holm & Sander)

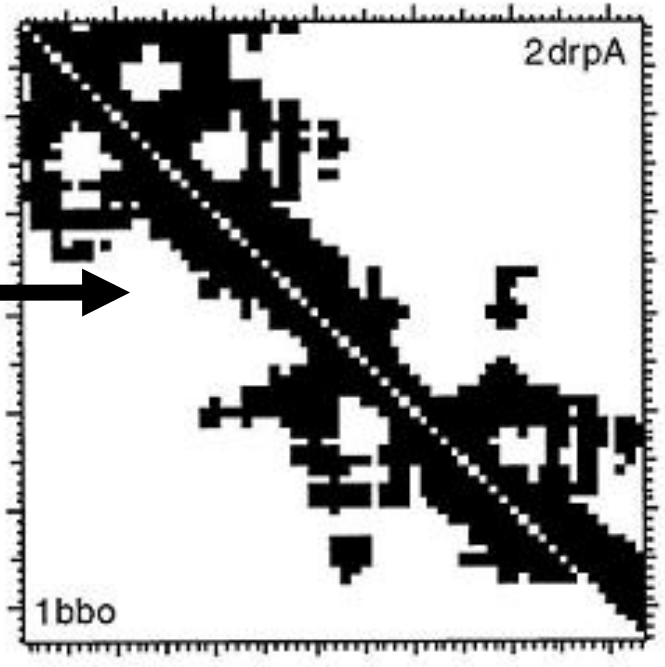


	1	2	3	4	5	6
1		3,2	5,6	3,9	3,5	
2			4,8	3,5	7,0	
3				3,6	6,7	
4					5,9	
5						
6						

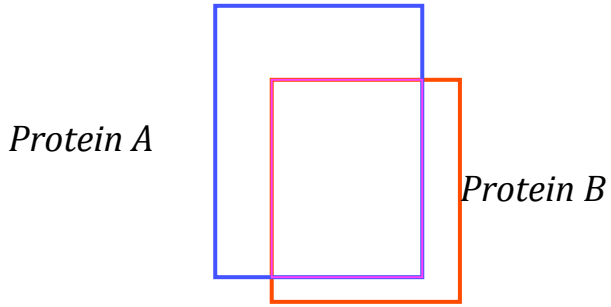
Each conformation is characterized by a Cα-Cα distance matrix d_{ij} (in Å)

Distance matrix (triangular) for the 2 proteins 2drp et 1bbo (black: contact – white: non contact)

DALI proceeds by comparison of the distance matrices - less accurate but much faster than rms – DALI can compare one structure to all structures in the PDB



Protein structure comparisons

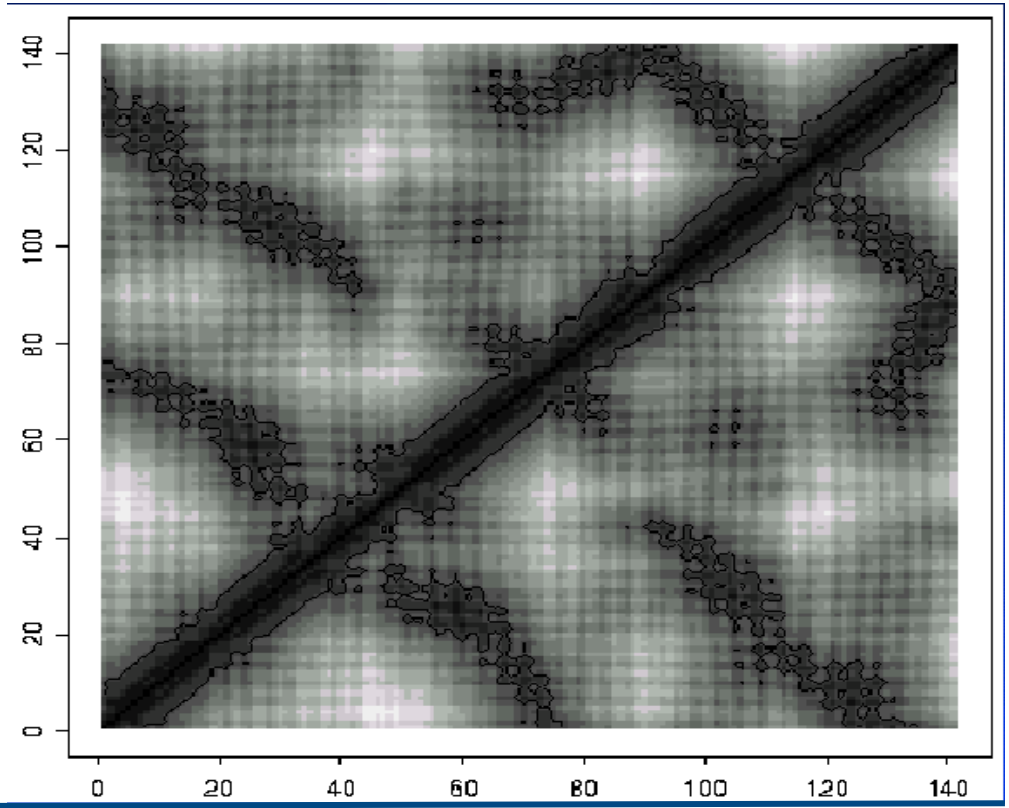


Principle: sliding the submatrices of one protein over the submatrices of the other protein - search for similarities

In practice:

- Matrices are divided into submatrices of fixed size (here hexapeptides)
- Comparing the submatrices two by two - one of each protein
- Assembling the pairs of submatrices for global alignment

Distance matrix of myoglobin



Protein structure comparisons

How to compare the different submatrices? Measure of structural similarity

$$Score = \sum_{i \in core} \sum_{j \in core} (\Theta - \Delta(d_{ij}^{A}, d_{ij}^{B}))\ \omega(d_{ij}^{A}, d_{ij}^{B})$$

- 'core' corresponds to the set of equivalent residues in proteins A and B
- Δ corresponds to the deviation of the distances d_{ij}^{A} and d_{ij}^{B} with respect to their arithmetic mean: $\Delta = \left| d_{ij}^{A} - d_{ij}^{B} \right| / \bar{d}_{ij}$
- Θ is a similarity threshold, determined empirically : $\Theta = 0.2$ (to ensure that known structural similarities are recovered
- $\omega = \exp(\bar{d}_{ij}^{2} / r^{2})$ with r=20Å => gives less weight to longer distances ; 20 Å is the typical size of a protein domain

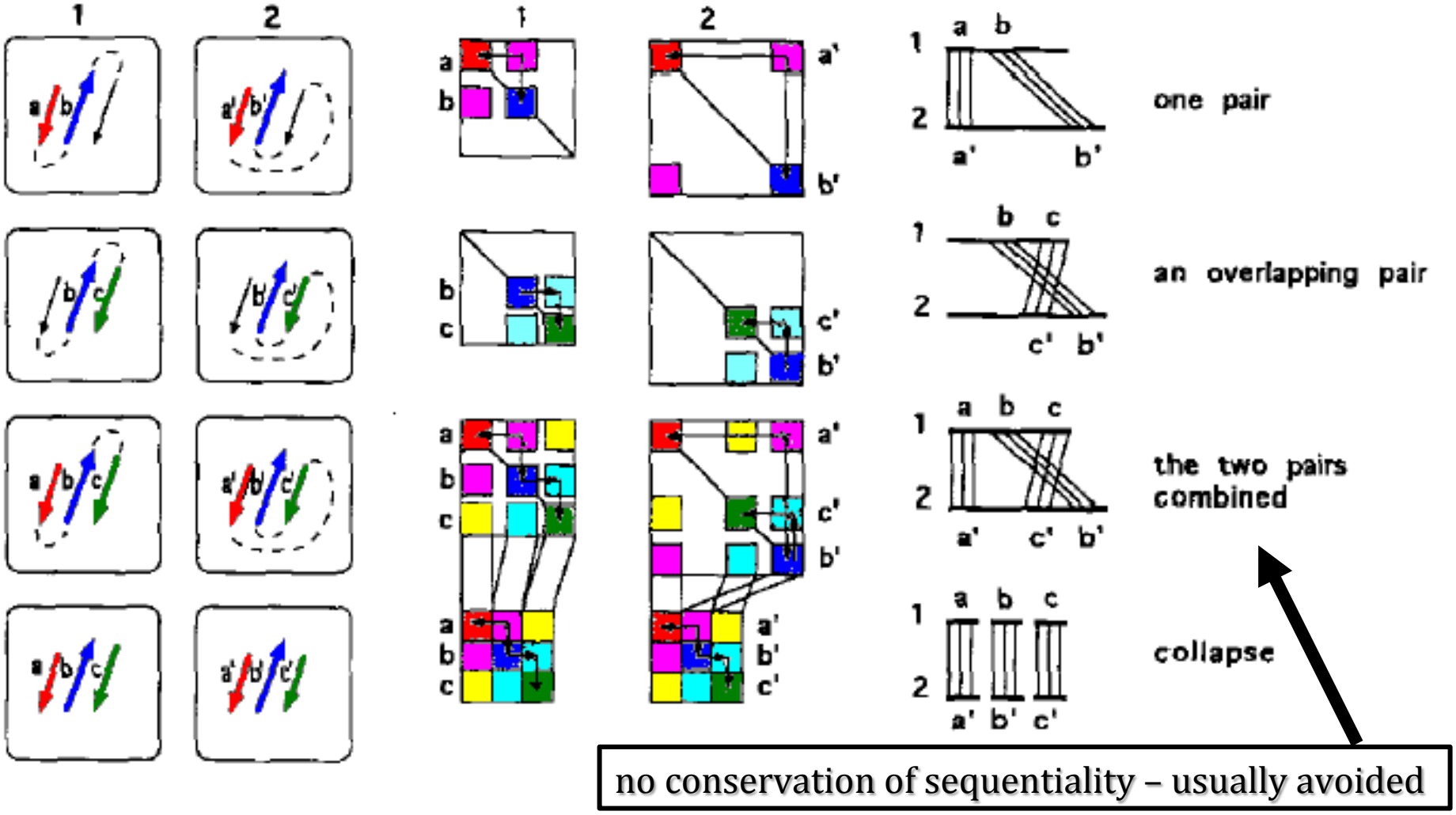
How to assemble the different submatrices ?

-> Non trivial optimization problem

Requires specific and fast algorithms (cf previous alignment method)

Algorithms used: branch & bound and Monte Carlo (see later)

Schematic view of how DALI functions



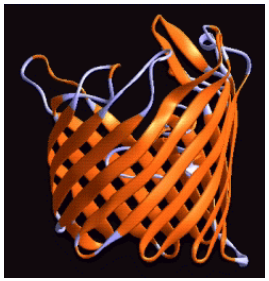
Protein structure classification

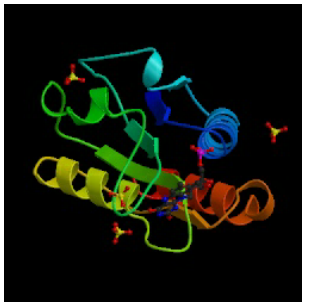
When analyzing protein structures, we see that structural motifs - some spatial organizations - are found in proteins with different sequences.

Structure classification plays an important role in understanding the structure, function and evolution of proteins.

Several levels of similarity

1) Secondary structures : α-helices, 3_{10}-helices, β-sheets, turns, loops, coil,...



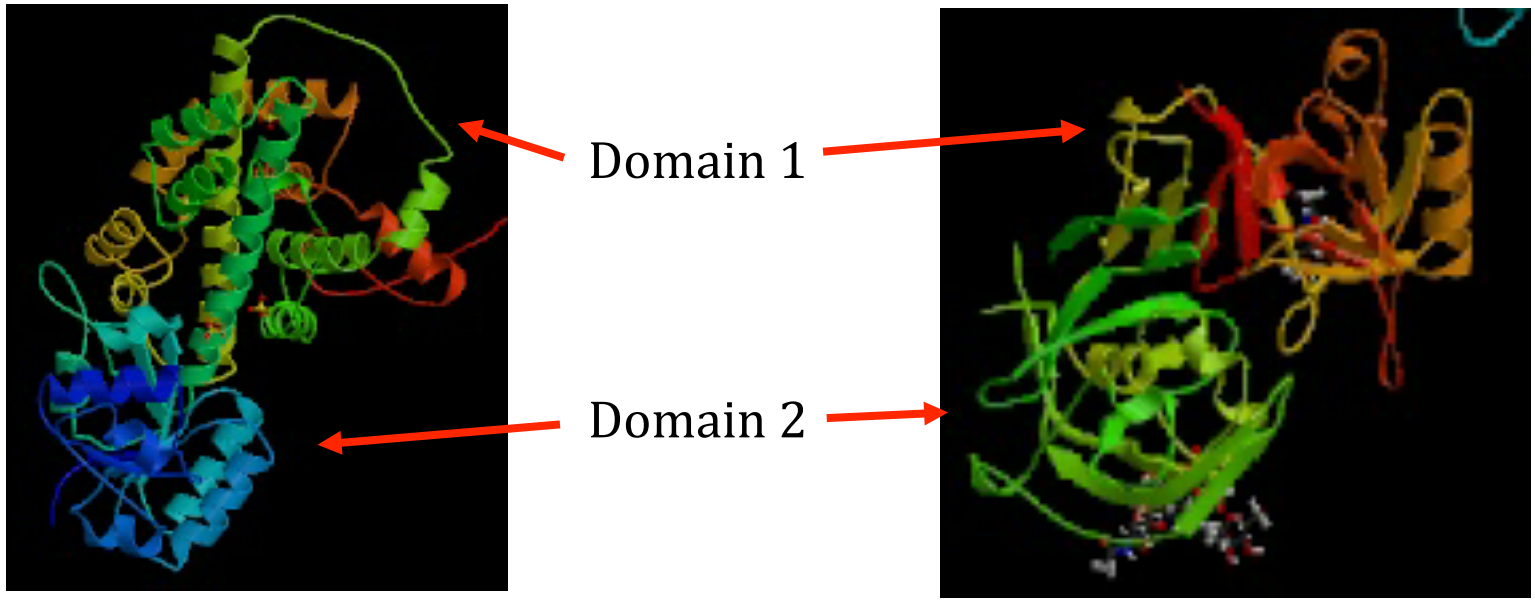


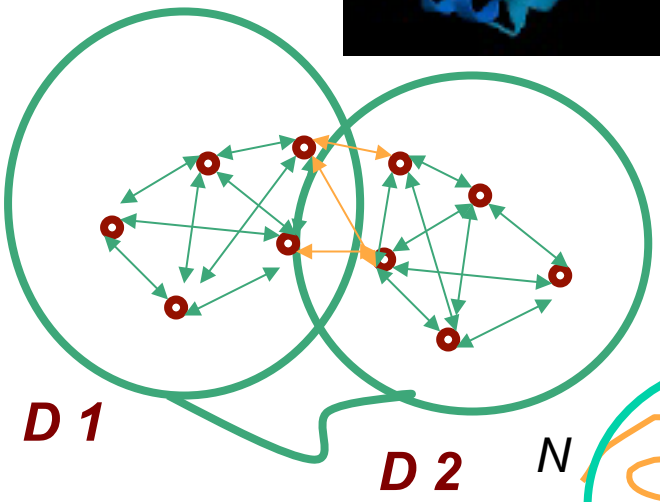
All α proteins, all β proteins, $\alpha+\beta$, ...

Protein structure classification -> domain classification

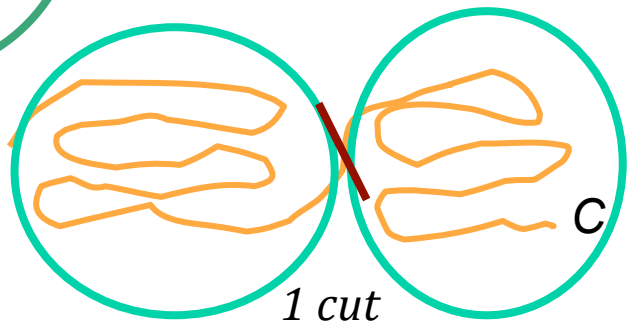
Often, proteins contain several domains =>

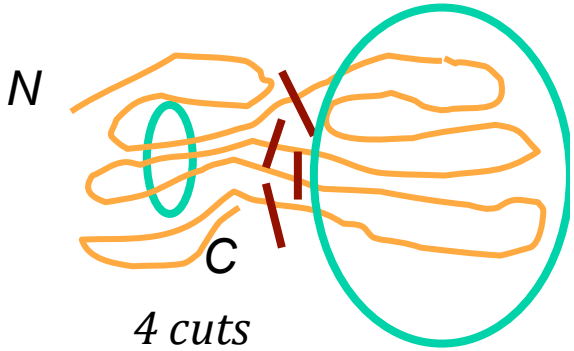
2) <u>Assignment into domains/ folding units</u>





Principle: More interactions among residues inside a domain than between domains => identification of groups of residues such that the number of contacts between groups is minimal





Protein structure classification

3) Assignment of a domain to a structural class:

- All α or all β : domains with predominantly α-helices or β-sheets – involves some subjectivity, where is the limit?

- α/β or α+β : domains with a mixture of α and β.
  α/β: intimate association of α and β, or contain many βαβ units (with adjacent β's)
  α+β: separate regions of α and β.
  Limit between α/β or α+β unclear => some classifications do not make the difference

- Other classes with few secondary structures - often small proteins whose tertiary structure is dominated by disulphide bridges, or one or more metal binding sites .

4) Fold assignment

Defined by the number, type, connectivity and arrangement of secondary structures - using structure alignment programs or by visual inspection.
! If two proteins differ by circular permutation (connect N- and C-termini and cut elsewhere), the similarity is often not detected.

Protein structure classification

5) Superfamilies

= Groups of proteins that appear homologous, even in the absence of significant sequence similarity.

Do all the proteins that have the same fold have a common ancestor? Impossible to answer these questions with certainty.
There exist convergent evolution (-> analogues) and divergent evolution (-> homologues). Argument in favor of the existence of homologues and analogues: the sequence identity between protein sequence pairs shows a bimodal distribution

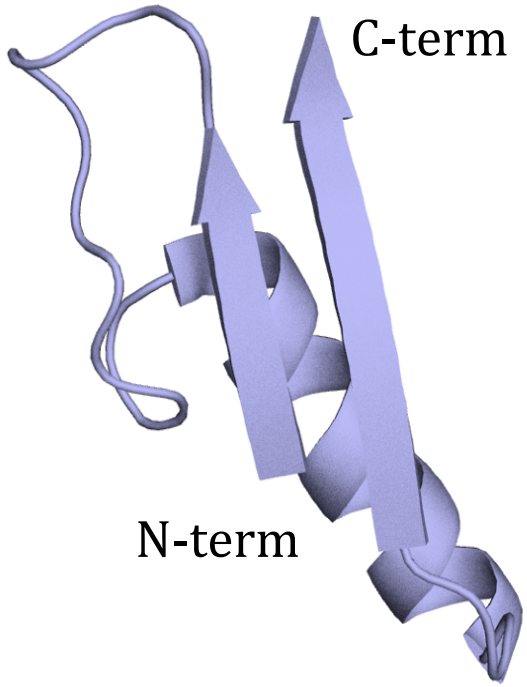
How can we distinguish between homologues and analogues?

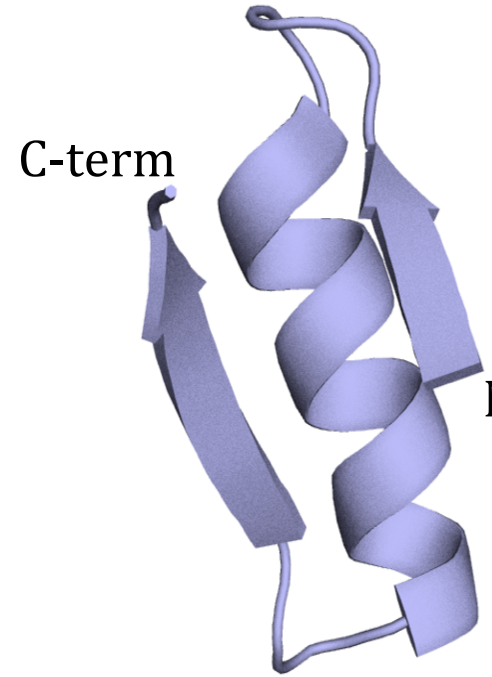
Two proteins are considered as homologoous :

- above a sequence similarity level.
- above a structure similarity level, even if the sequence similarity is insignificant.
- If unusual structural features are conserved: conformations of functionally important turns/loops, left βαβ units (see next slide)
- low but significant sequence identity detected upon alignment of the structures.
- conservation of key residues in the active site, even in the absence of overall sequence similarity
- by transitivity, A and B are homologous, B and C also => A and C also (dangerous!)

Protein structure classification

Left-handed βαβ unit
(uncommon)





Right-handed βαβ unit
(common)
(note: helix is under the β-strands)

Protein structure classification

6) Superfolds, supersites

Superfolds: some folds are populated by different superfamilies

-> suggests that the folding has appeared several times by convergent evolution.

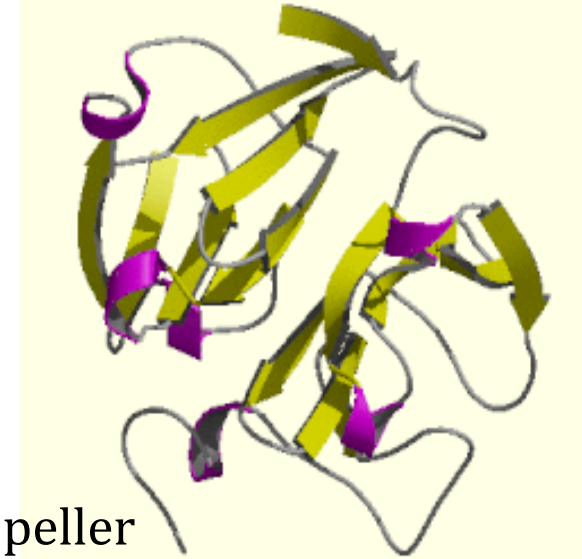
Examples : α/β barrel (or TIM-barrel), β-propellers, …

Proteins belonging to the same superfold tend to bind ligands in the same sites: called supersites

These sites seem to be a property of the superfold, which dictates the best binding region, regardless of the evolutionary origin

α/β barrel

=> For some superfolds: possible to make predictions about the binding site even in the absence of information about a common ancestor.



β-propeller

Protein structure classification: Prediction of function from structural similarity

For ~ 50% of new structures the binding site can be accurately predicted from structure comparisons => Danger to interpret any structural similarity as an indication of a common function

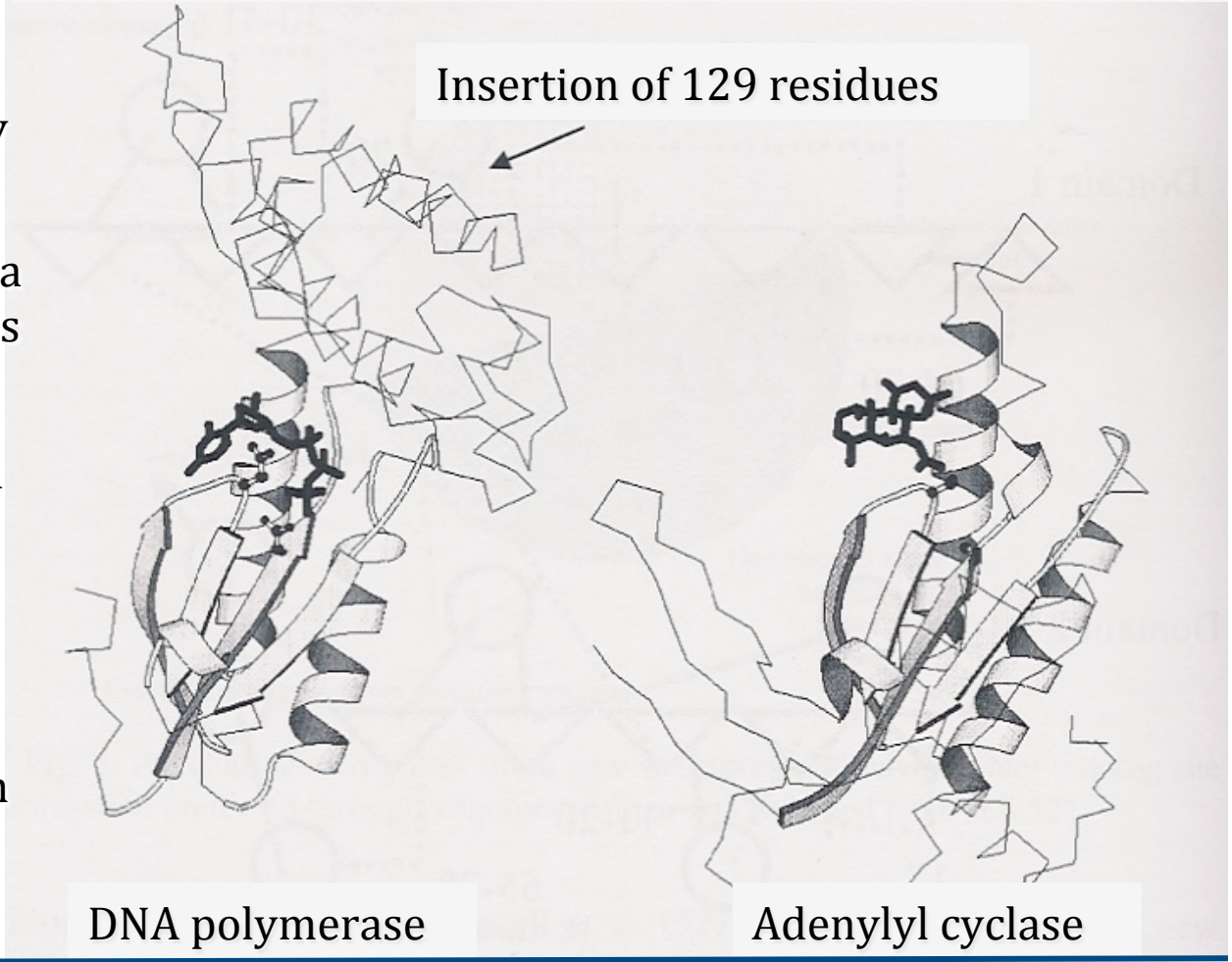
e.g. adenylyl cyclase: originally described as a new fold

But has clear similarities with a class of RNA/DNA polymerases

The cyclases and polymerases have a similar binding site and reaction mechanism, and require Mg^{2+} ions binding to Asp

=> Structural similarity gives information on the mechanism of action of the little known cyclases.

The adenylyl cyclase synthesizes cyclic AMP from ATP



Protein structure classification:

2 classifications available on the web

CATH

C (class)

1. all α
2. all β
3. α and β
4. small proteins with few secondary structures

A (architecture)

General shape of the structure, taking into account the relative orientation of the secondary structures and ignoring the connectivity

T (topology)

Structures grouped according to the general shape and the connectivity of the secondary structures.

H (homologous superfamily)

Homologous protein domains

Also: **Sequence families**

Structures grouped as a function of their sequence identity (≥ 35%).

SCOP

Class

More subdivisions than in CATH + classes for membrane proteins, small proteins, ...

Fold

Equivalent to **T**opology of CATH

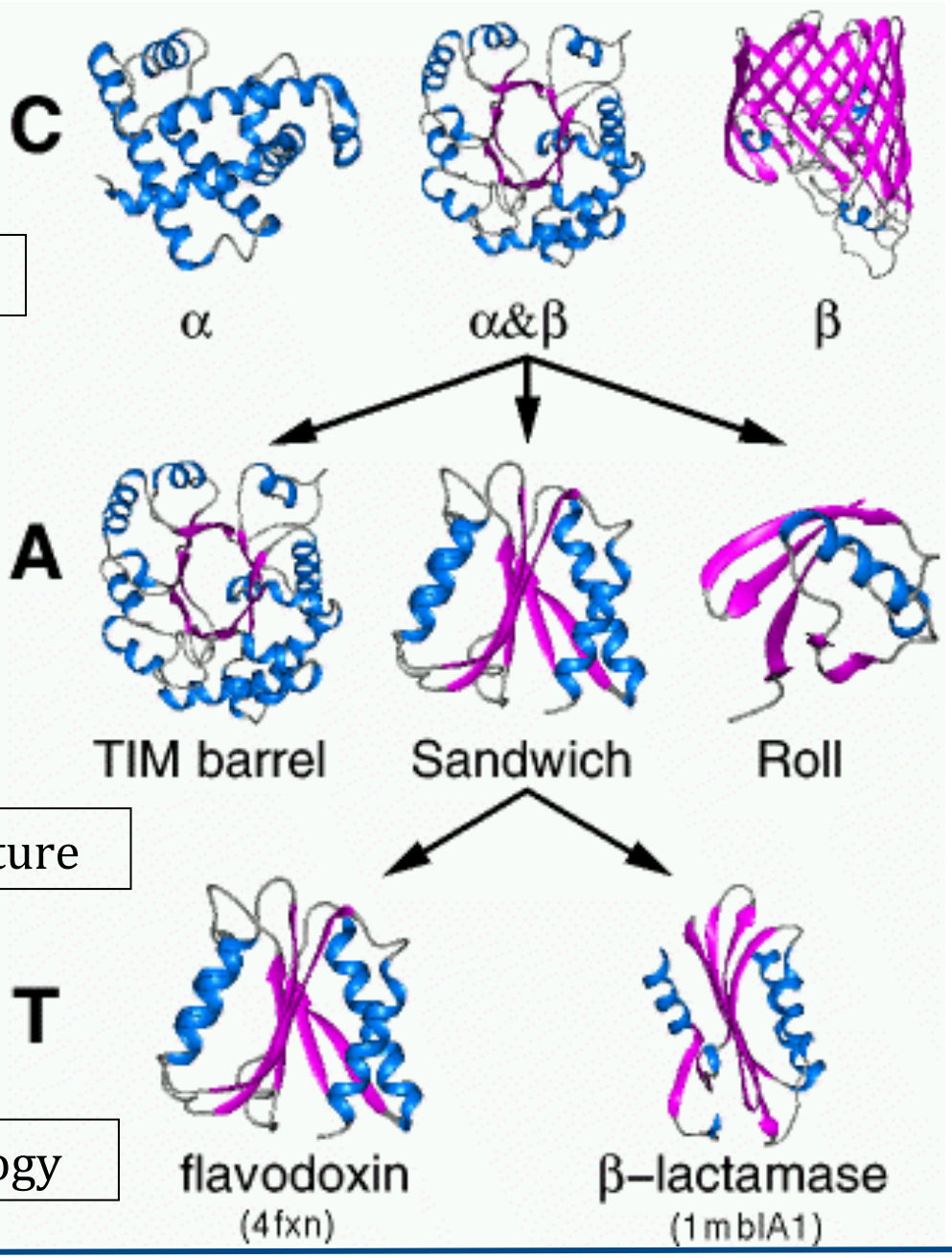
Superfamily

Equivalent to **H**omologous superfamily of CATH

Family

Equivalent to Sequence families of CATH, but without the precise threshold of 35%.

Protein structure classification



Class

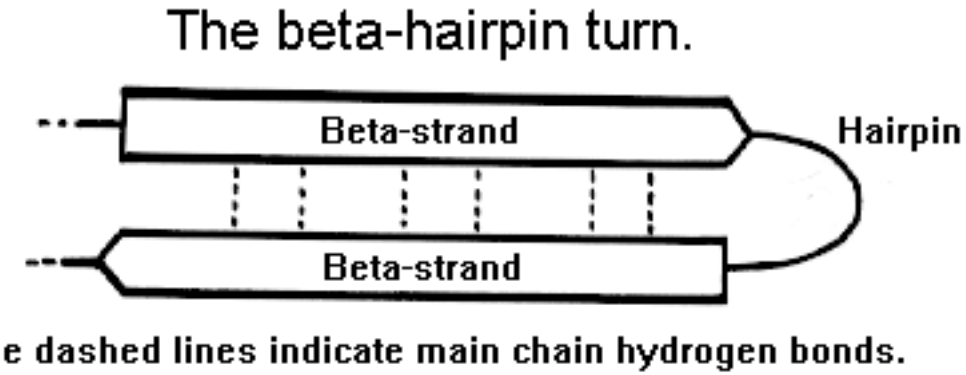
Architecture

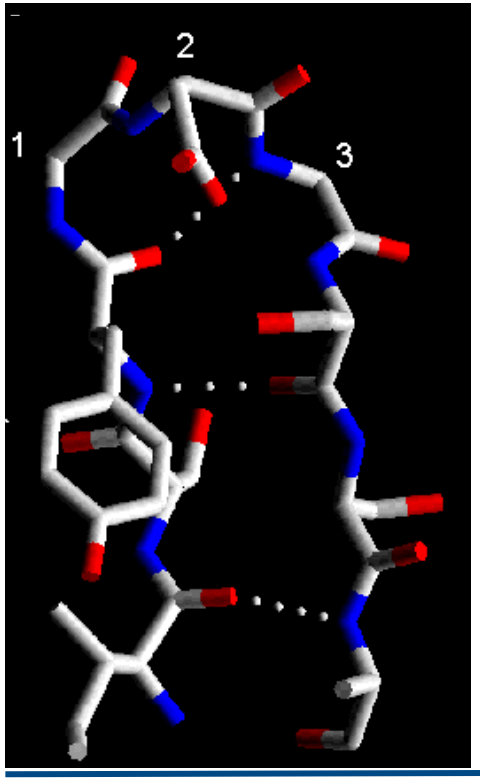
Topology

Example from CATH

Protein structure classification

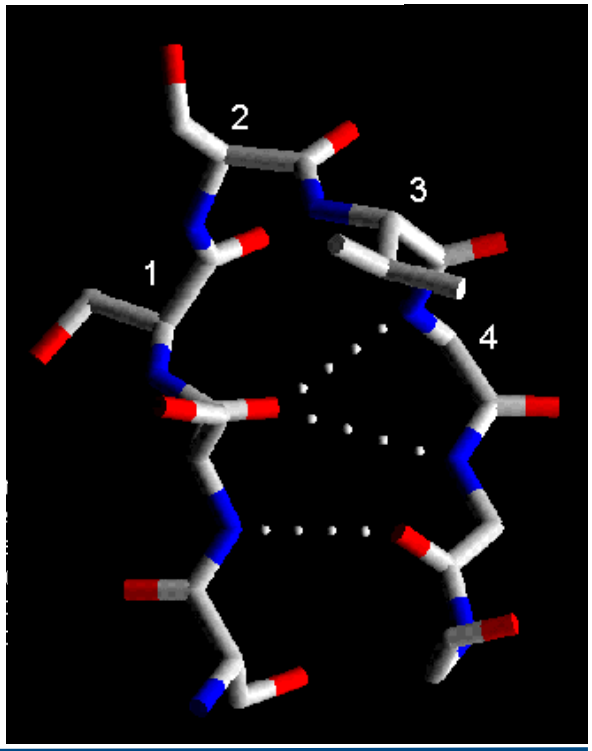
Classification of recurrent motifs in proteins
Example of classes of turn motifs





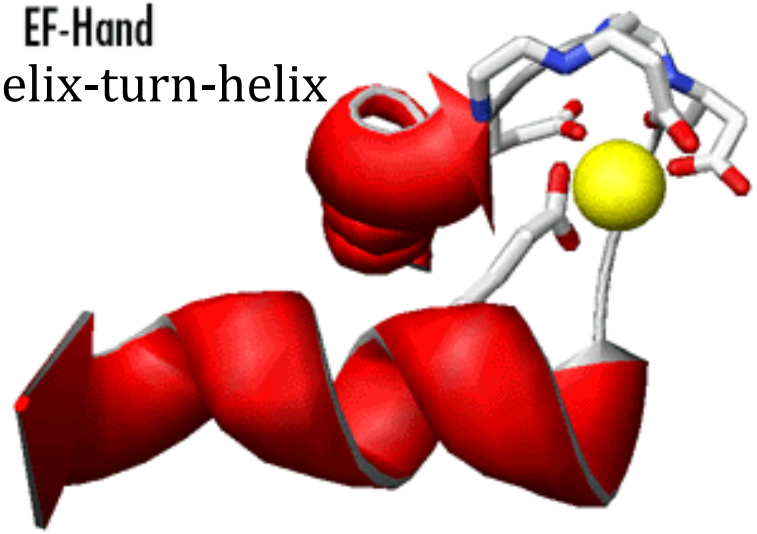
3-residue β-hairpin

4-residue β-hairpin

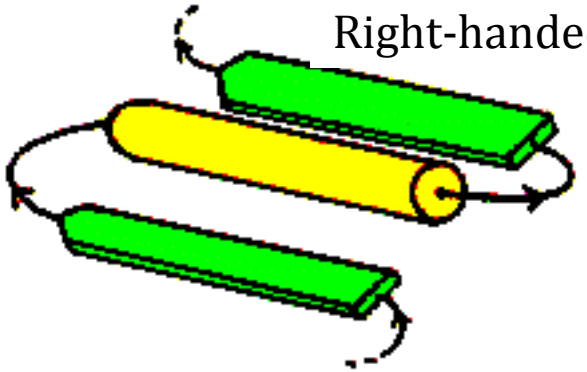


Protein structure classification

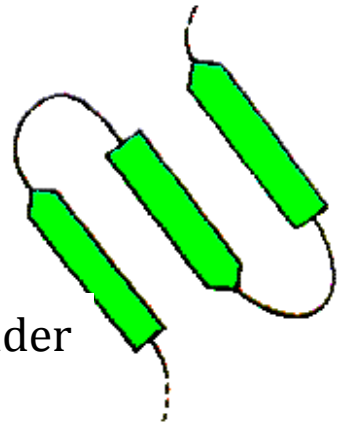
EF-Hand
Helix-turn-helix



Right-handed βαβ unit



β meander



Other types of turn motifs