UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL ATOMIC ENERGY AGENCY
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

H4.SMR/916 - 25

SEVENTH COLLEGE ON BIOPHYSICS:

Structure and Function of Biopolymers: Experimental and Theoretical Techniques.

4 - 29 March 1996

DNA-Protein Interactions

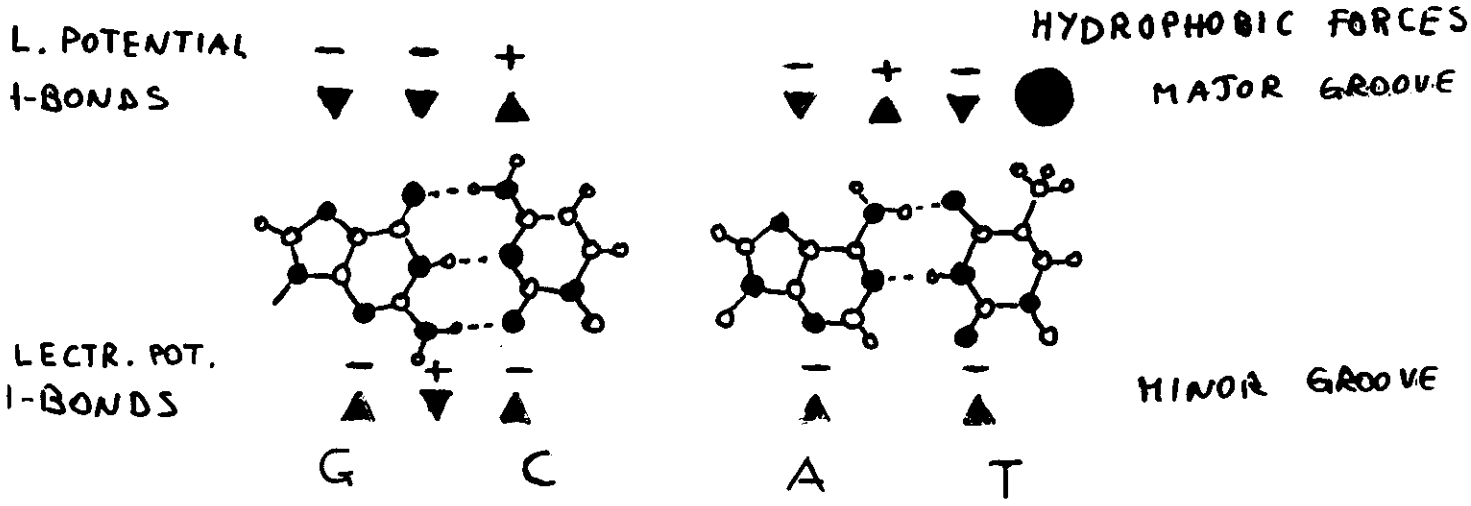
Franco QUADRIFOGLIO
Dept. Biomedical Sciences and Technologies
School of Medicine
University of Udine
Udine

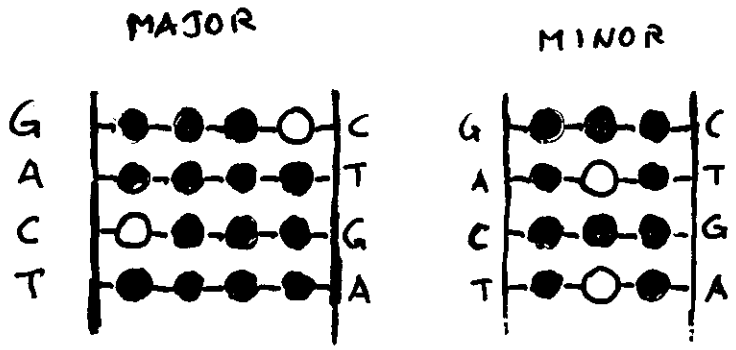
Franco Quadrifoglio
Dept. Biomedical Sciences and Technologies
School of Medicine
University of Udine

DNA-Protein Interactions

(Lecture Notes)

DNA RECOGNITION BY PROTEINS

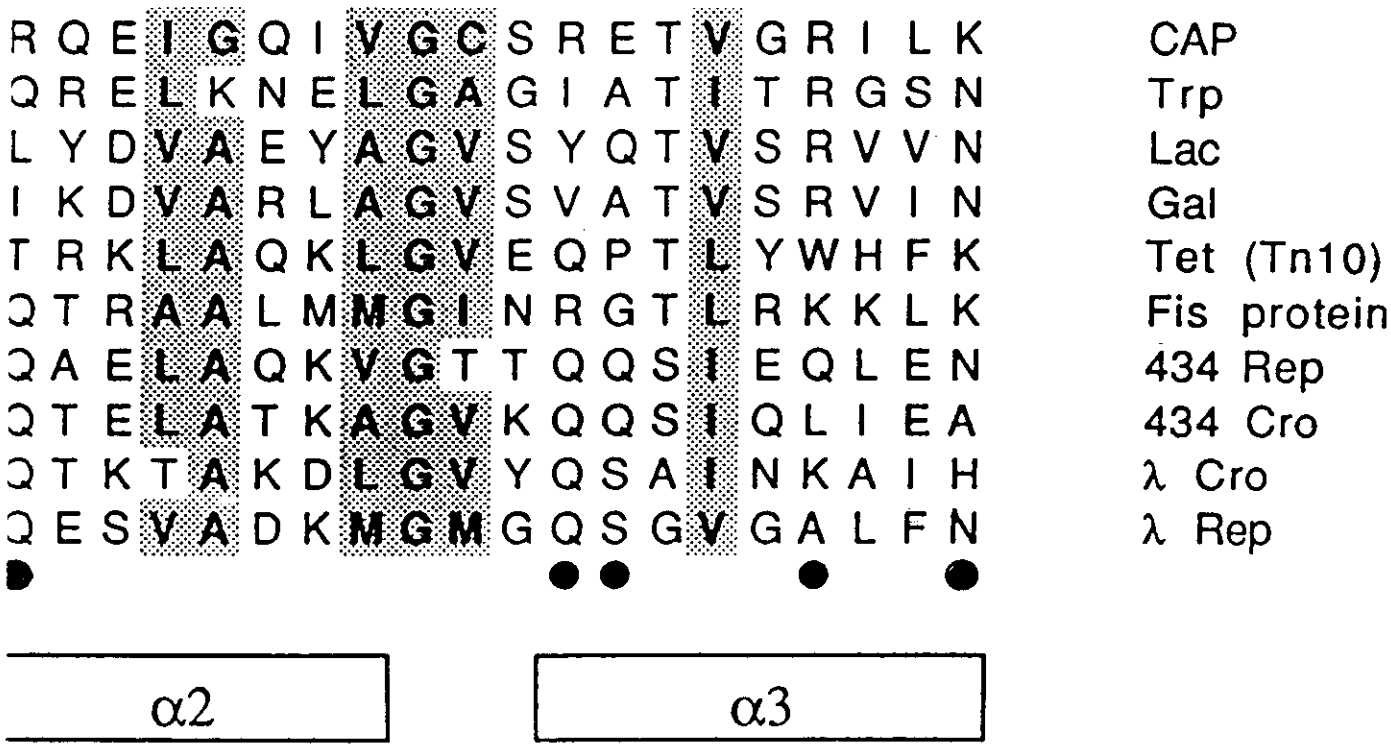




● ACCEPTOR
● DONOR
○ HYDROGEN ATOM
● METHYL GROUP

Fig. 19

HTH MOTIF



HYDROPHOBIC OR SMALL RESIDUES ARE SHADED

● BACKBONE CONTACTS

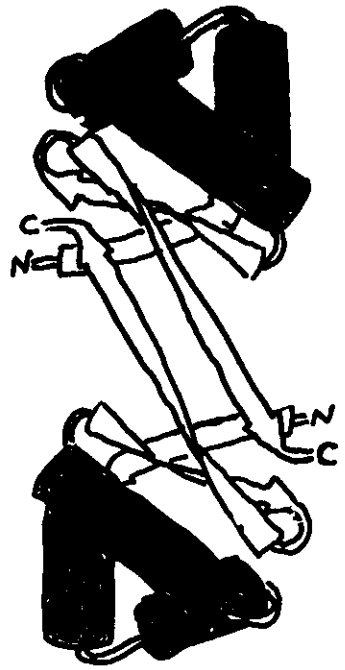
● HYDROGEN BONDS WITH BASES IN THE MAJOR GROOVE

COMMON FEATURES:

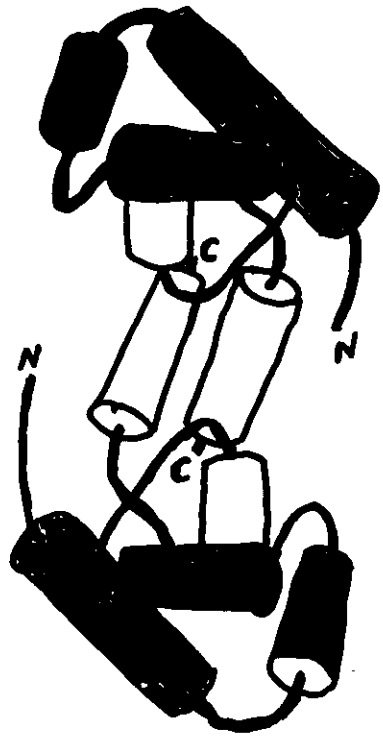
- THE PROTEINS BIND AS DIMERS
- THE FIRST HELIX IN THE HTH UNIT IS "ABOVE" THE MAJOR GROOVE BUT ITS N-TERMINAL PORTION CONTACTS THE DNA BACKBONE.
- THE SECOND HELIX FITS THE MAJOR GROOVE AND ITS N-TERMINAL PORTION "SEES" THE EDGES OF THE BASE-PAIRS
- DNA IS IN B-FORM
- EXTENSIVE NETWORK OF HYDROGEN BONDS BETWEEN THE PROTEIN AND THE DNA BACKBONE.

FIG. 2

HTH



LAMBDA CRO PROTEIN



LAMBDA REPRESSOR

SEQUENCE RECOGNITION OF AN INTERRUPTED PALINDROMIC SEQUENCE IS THE MAJOR DETERMINANT IN PROTEIN BINDING. THIS HAS BEEN DEMONSTRATED BY GENETIC ANALYSIS AND SWAP EXPERIMENTS ON RECOGNITION HELIX.

HOWEVER ADDITIONAL CONTACTS OUTSIDE HTH REGION CONTRIBUTE TO THE AFFINITY OF THE PROTEIN FOR THE OPERATOR SITE. THE N-TERMINAL REGION, FOR EXAMPLE, MAKES CONTACTS WITH THE OPPOSITE SIDE OF THE DNA SURFACE, WHEREAS MANY CONTACTS WITH BACKBONE PHOSPHATES ARE PRESENT.

FIG. 3

λ REPRESSOR - DNA COMPLEX

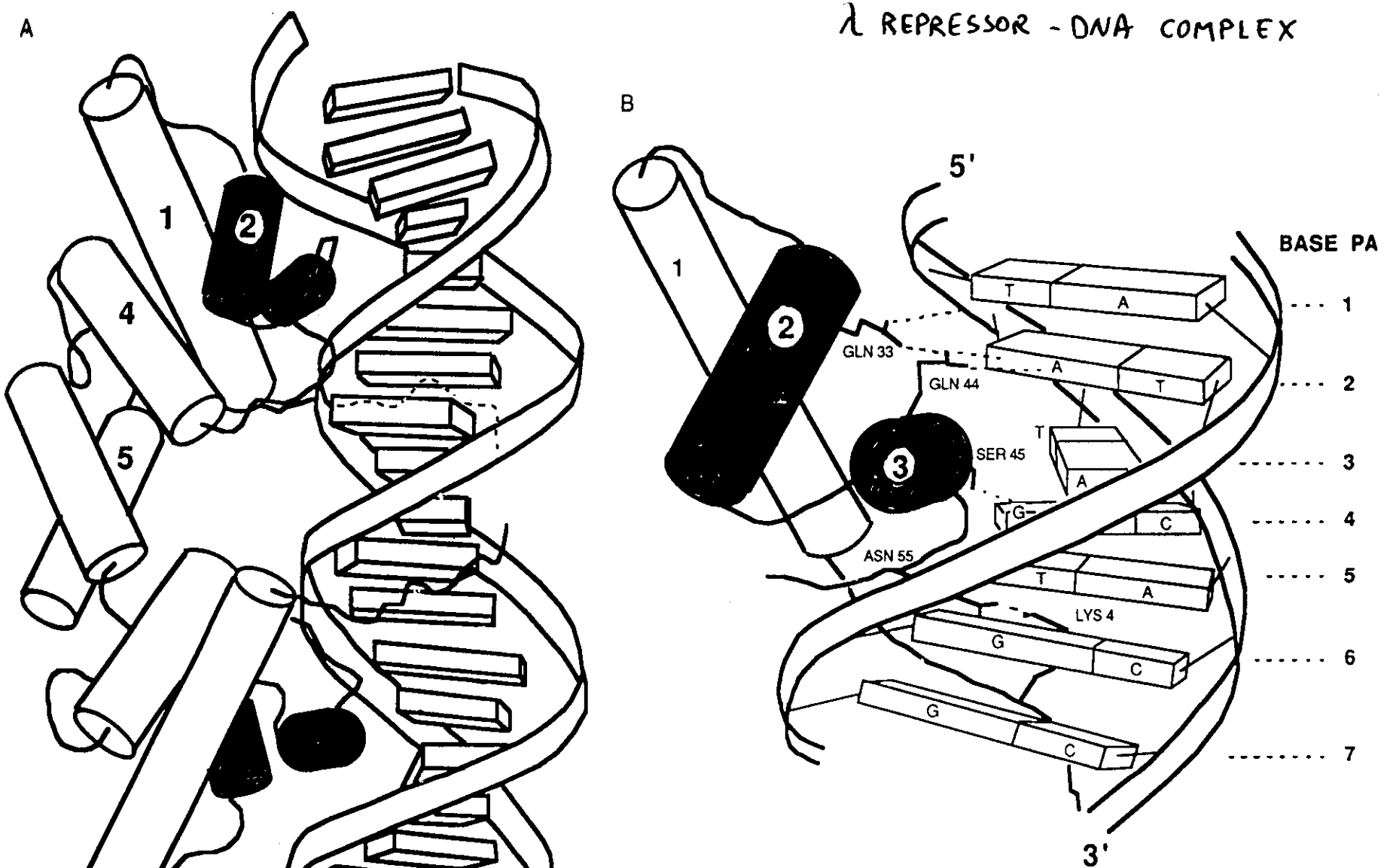
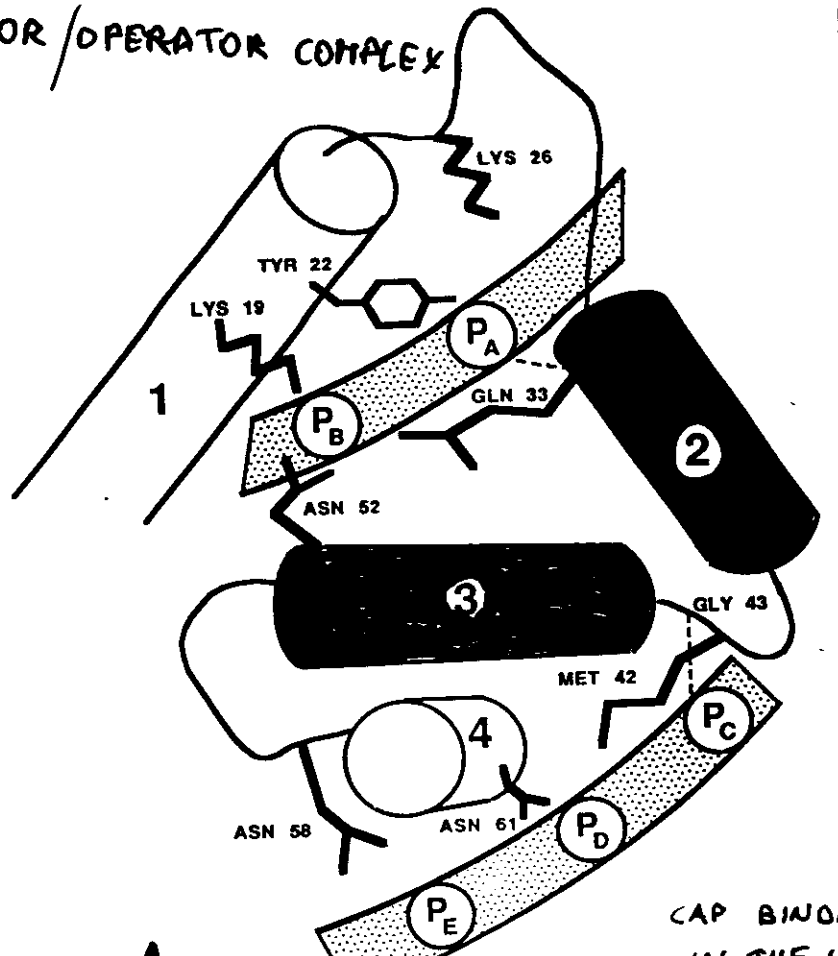


FIG. 4

λ-REPRESSOR/OPERATOR COMPLEX

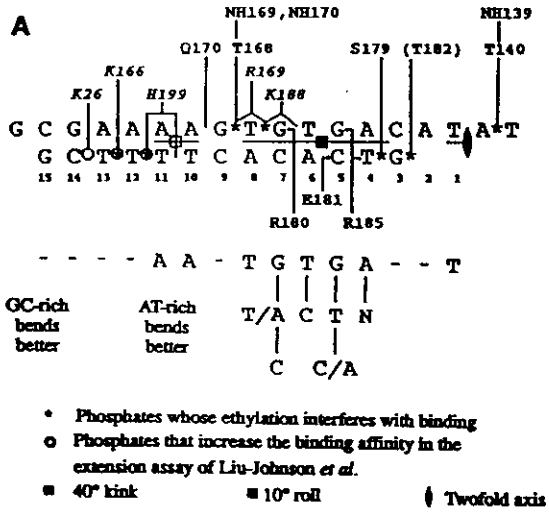


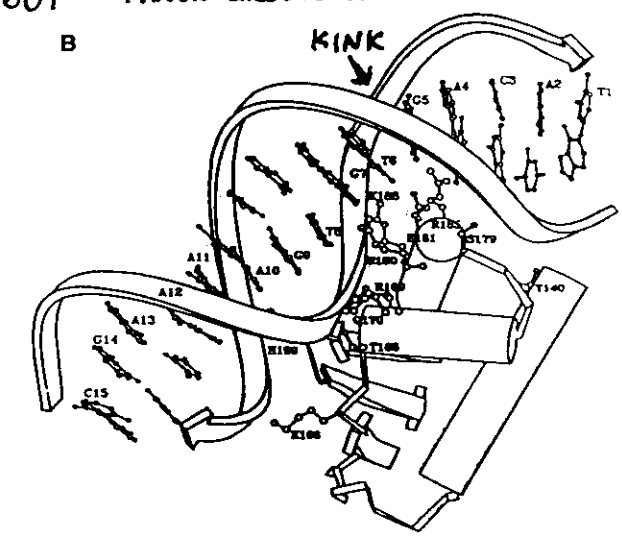
CAP BINDING SEQUENCE IN THE LAC OPERON

AACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACC

OUT IN OUT IN OUT IN OUT

MINOR GROOVE ORIENTATION





cAMP-CAP/DNA COMPLEX

Schultz et al Science 253, 1001 (1991)

FIG. 5

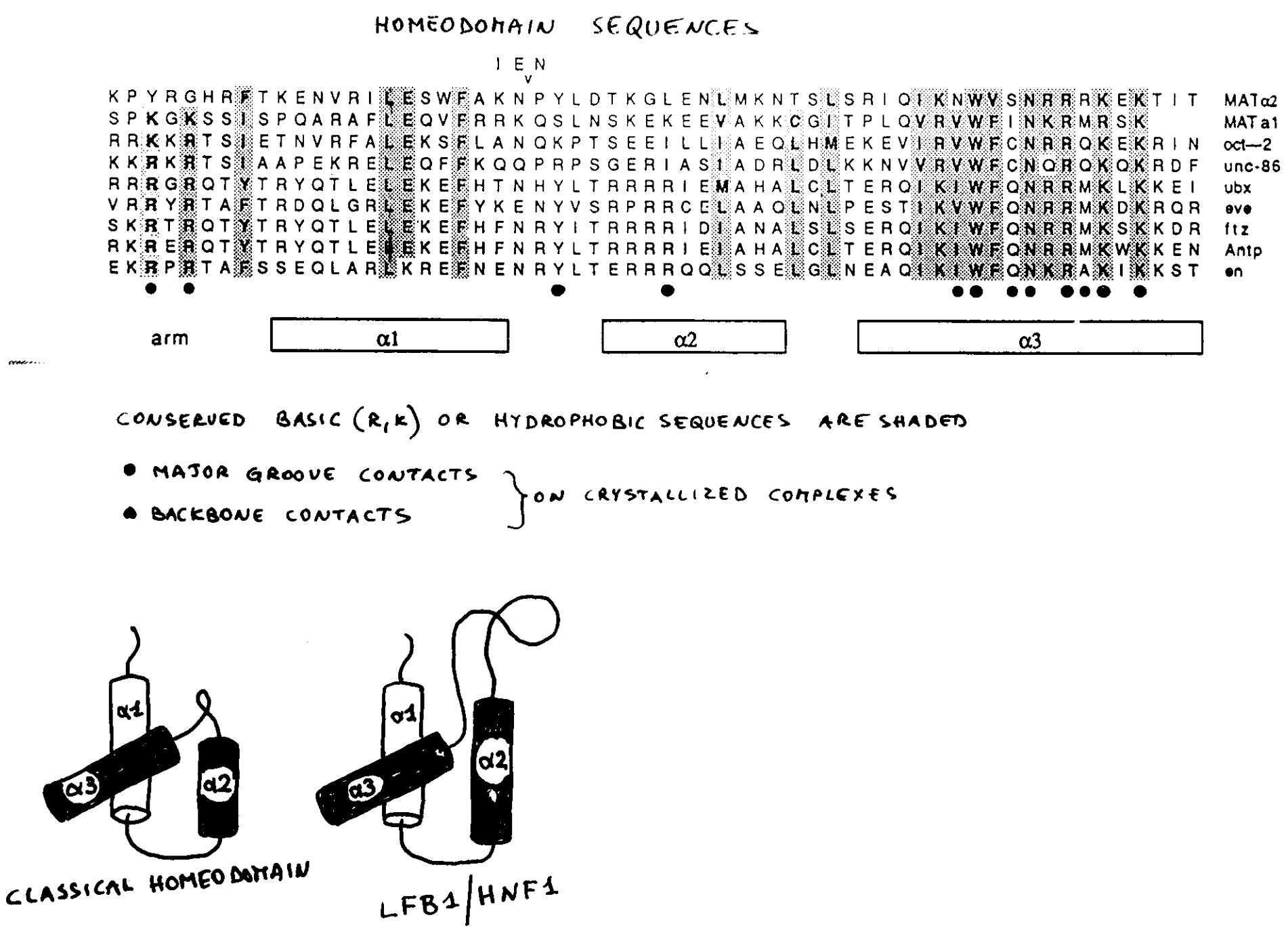

HOMEODOMAIN SEQUENCES
I E N
K P Y R G H R F T K E N V R I L E S W F A K N P Y L D T K G L E N L M K N T S L S R I Q I K N W V S N R R R K E K T I T MATα2
S P K G K S S I S P Q A R A F L E Q V F R R K Q S L N S K E K E E V A K K C G I T P L Q V R V W F I N K R M R S K MATa1
R R K K R T S I E T N V R F A L E K S F L A N Q K P T S E E I L L I A E Q L H M E K E V I R V W F C N R R Q K E K R I N oct—2
K K R K R T S I A A P E K R E L E Q F F K Q Q P R P S G E R I A S I A D R L D L K K N V V R V W F C N Q R Q K Q K R D F unc-86
R R R G R Q T Y T R Y Q T L E L E K E F H T N H Y L T R R R R I E M A H A L C L T E R Q I K I W F Q N R R M K L K K E I ubx
V R R Y R T A F T R D Q L G R L E K E F Y K E N Y V S R P R R C E L A A Q L N L P E S T I K V W F Q N R R M K D K R Q R eve
S K R T R Q T Y T R Y Q T L E L E K E F H F N R Y I T R R R R I D I A N A L S L S E R Q I K I W F Q N R R M K S K K D R ftz
R K R E R Q T Y T R Y Q T L E L E K E F H F N R Y L T R R R R I E I A H A L C L T E R Q I K I W F Q N R R M K W K K E N Antp
E K R P R T A F S S E Q L A R L K R E F N E N R Y L T E R R R Q Q L S S E L G L N E A Q I K I W F Q N K R A K I K K S T en
arm
α1
α2
α3
CONSERVED BASIC (R,K) OR HYDROPHOBIC SEQUENCES ARE SHADED
MAJOR GROOVE CONTACTS
BACKBONE CONTACTS
} ON CRYSTALLIZED COMPLEXES
α1
α2
α3
CLASSICAL HOMEODOMAIN
α1
α2
α3
LFB1/HNF1

FIG. 6

HOMEODOMAIN

- 60 CONSERVED RESIDUES WHICH FORM A STABLE, FOLDED STRUCTURE, ABLE TO BIND DNA BY ITSELF AND WHICH CONTAINS A HTH MOTIF.

- THE HOMEODOMAIN COMPRISES AN EXTENDED N-TERMINAL ARM AND THREE α-HELICES

- HELIX 1 AND HELIX 2 PACK AGAINST EACH OTHER. HELIX 3 IS ALMOST PERPENDICULAR AND ITS HYDROPHOBIC SIDE PACKS AGAINST HELICES 2 AND 1

- HELIX 3 IS MAINLY RESPONSIBLE OF THE CONTACTS WITH DNA IN THE MAJOR GROOVE, ESPECIALLY WITH ITS CENTRAL REGION

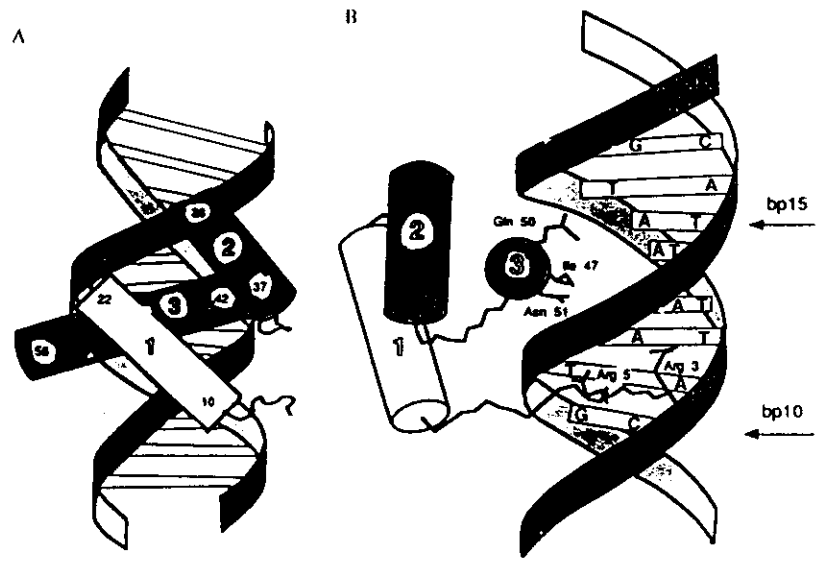
- OTHER CONTACTS ARE MADE BY HELICES 1 AND 2 WITH DNA BACKBONE BUT PRIMARILY BY N-TERMINAL ARM WITH THE MINOR GROOVE

- INVARIANT RESIDUES IN THE HELIX 3 ARE NECESSARY FOR THE CORRECT FOLDING (Trp 48 AND Phe 49), OR FOR CRITICAL CONTACTS WITH DNA (Asn 51 AND Arg 53).

- THE HTH OF HOMEODOMAIN IS DIFFERENT FROM THE CLASSICAL HTH OF PROCARYOTES. HELIX 3 IN THE HOMEODOMAINS CONTACTS DNA IN THE CENTER, WHEREAS PRO. HTH CONTACTS DNA WITH THE N-TERMINAL PORTION OF THE HELIX - HELIX 2 IN THE PROCARYOTIC PROTEINS MAKES DNA CONTACTS

A NON CLASSICAL HOMEODOMAIN HAS BEEN FOUND ON RAT LIVER TF LFB1/HNF1. THREE HELICES ARE ORIENTED AND PACKED AS ANTP, EN AND MATα2 HOMEODOMAIN, BUT LFB1/HNF1 CONTAINS A 21 RESIDUE INSERTION BETWEEN HELICES 2 AND 3 THIS INSERTION COULD MEDIATE DIMERIZATION OF LFB1/HNF1 ON AN INVERTED PALINDROME CONSENSUS SEQUENCE.

FIG. 7



ENGRAILED HD /DNA COMPLEX

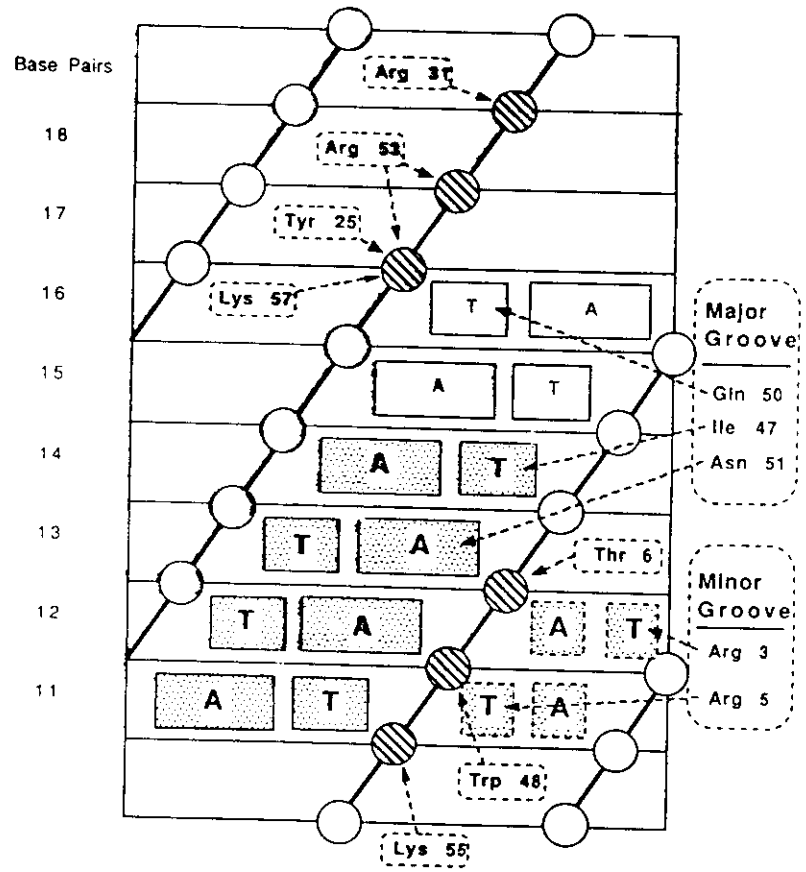


FIG. 8

POU DOMAINS (Pit-1, Oct-1, Oct-2, Unc-86)

IT IS A BIPARTITE DNA BINDING DOMAIN CONSISTING OF

A) A POU-SPECIFIC (POU_S) DOMAIN
B) A POU-HOMEODOMAIN (POU_H) } CONNECTED BY A LINKER

THE POU_H IS SIMILAR TO OTHER HOMEODOMAINS

POU_S CONSISTS OF FOUR HELICES PACKED TOGETHER BY A HYDROPHOBIC CORE.

IT IS VERY SIMILAR TO THE DNA-BINDING DOMAINS OF λ- AND 434-REPRESSORS AND [illegible] OF 434 CRO. THE FIFTH HELIX, WHICH IN PROKARYOTES, SERVES AS DIMERIZATION HELIX IS LACKING BECAUSE POU_S IS CONNECTED TO POU_H BY A LINKER.

HELICES 2 AND 3 REPRESENT HTH MOTIF, BUT THE TURN CONTAINS SIX EXTRA RESIDUES.

HELIX 3 RECOGNIZES THE OCTAMER MOTIF IN THE DNA (ATGCAAAT)

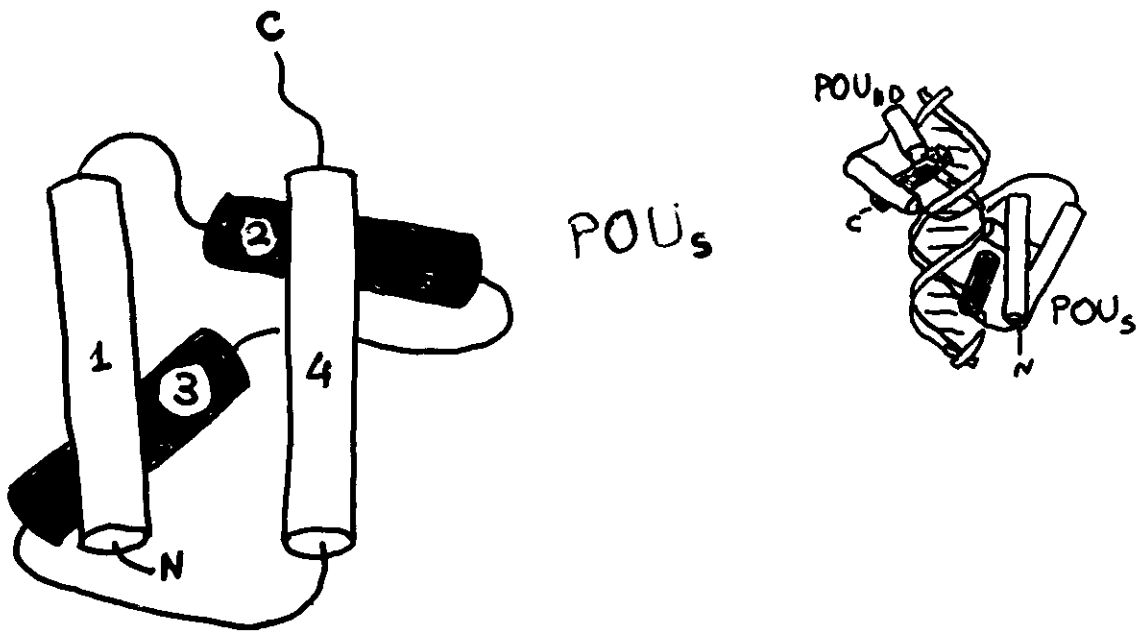


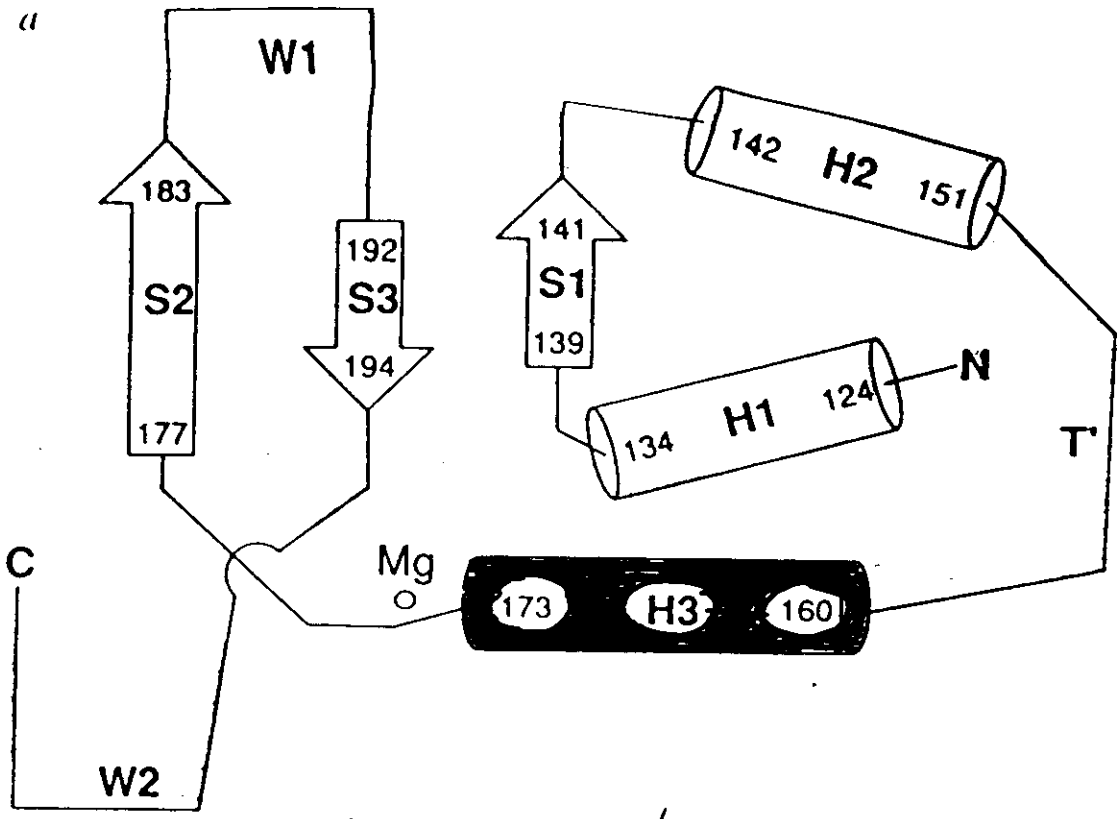
FIG. 9

HNF-3/ FORK HEAD

- FAMILY OF EUKARYOTE TF FOUND IN VARIOUS ORGANISMS FROM YEAST TO MAN
- MONOMERIC α/β PROTEIN STRUCTURE INTERACTING WITH B-DNA
- N-TERMINAL HALF HAS THREE α-HELICES WITH HELIX 3 INTERACTING WITH DNA MATOR GROOVE
- CARBOXY-TERMINAL HALF CONSISTS OF β-STRANDS AND RANDOM COIL AND MAKES VARIOUS DNA CONTACTS
- T' LOOP CONNECTING H2 AND H3 RUNS ALONG PHOSPHODIESTER BACKBONE
- THE N-TERMINAL HALF IS TOPOLOGICAL SIMILAR TO THE THREE α-HELICES OF HOMEODOMAIN BUT IS INTIMATELY ASSOCIATED WITH THE C-TERMIMAL HALF OF THE PROTEIN
- AN Mg^{+2} ION STABILIZES THE COMPACT STRUCTURE OF THE PROTEIN
- THE DNA STRUCTURE IS A SLIGHTLY BENT ($\sim 13°$) B-DNA WITH NARROWING OF THE MAJOR GROOVE
- THE STRUCTURE OF HNF-3γ CAN BE LIKENED TO A BUTTERFLY WITH AN α-HELICAL THORAX AND TWO WING-LIKE LOOPS AS TO BE CALLED "WINGED-HELIX" MOTIF
- THE STRUCTURE OF HNF-3 RESEMBLES THAT OFF THE HISTONE H5

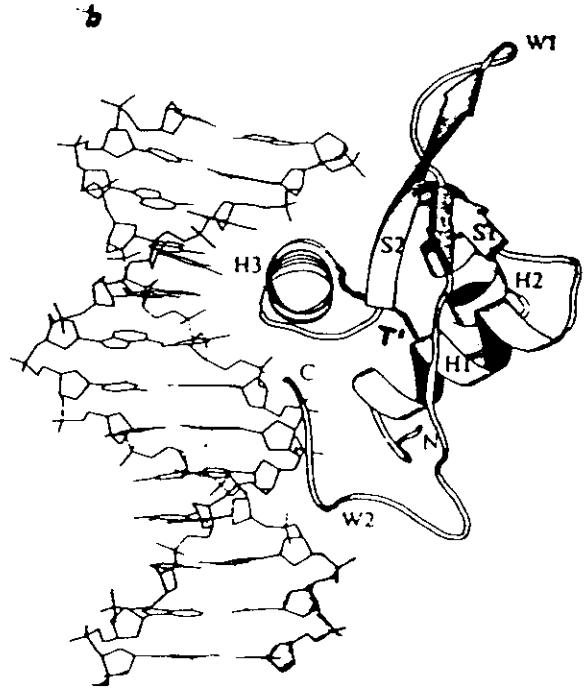
FIG.10

a



HNF-3γ/DNA COMPLEX

b



5' GACTAAGTCAACC 3'

FIG. 11

ZINC FINGERS DNA BINDING DOMAINS

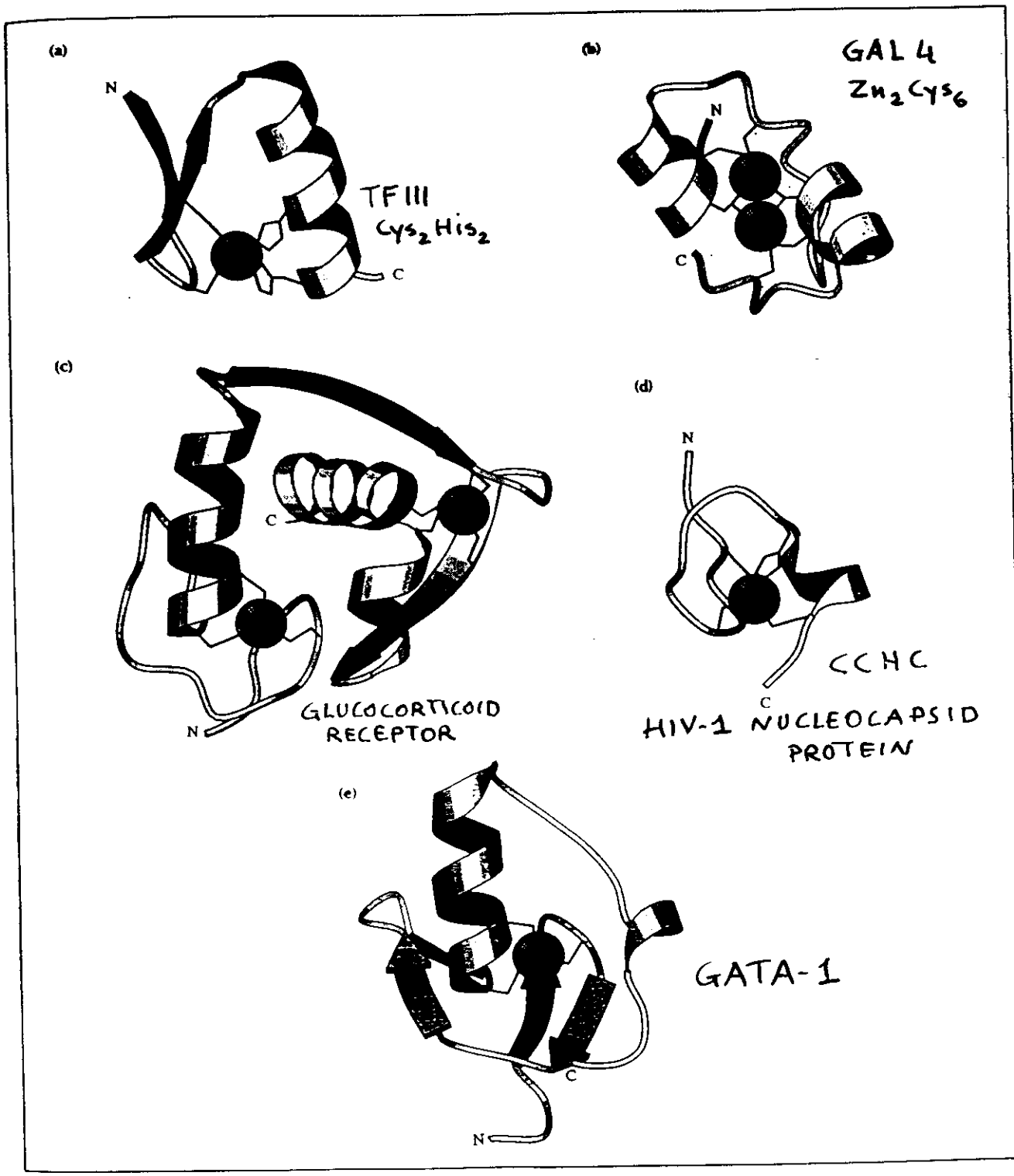


FIG. 14.1

ZINC FINGERS - C_2H_2

Proteins in this family contain tandem repeats of the 30-residue zinc finger motif

$$Cys - X_{2/4} - Cys - X_{12} - His - X_{3-5} His$$

The domain contains an antiparallel β-sheet and an α-helix. The coordinate zinc atom holds these structures to form a compact globular domain. The crystal structure of ZIF268/DNA shows that the proteins binds in the major groove with the α-helix. Each finger docks in the same way and makes contacts with a three base-pair subsite. The position of each finger is related to the other by a rotation of 96° (3 × 32°) and a translation of 10.2 Å (3 × 3.4 Å). One of the β-strand contacts the phosphate backbone.

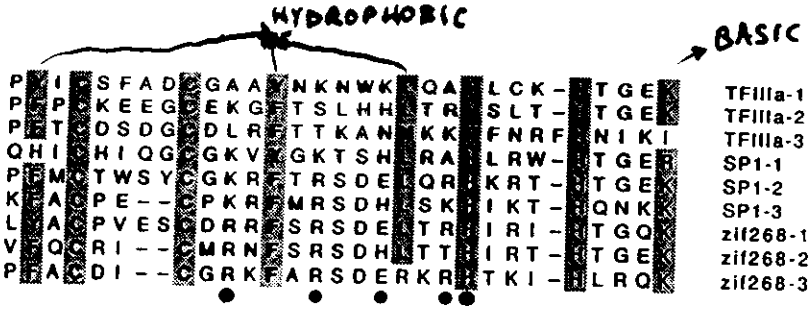
In each finger arginine residues make hydrogen bonds with guanine: the binding sequence is in fact rich of guanines. In general three contacts are made by each finger.

Recent examples of cocrystal structures show a third β-strand and different contacts with bases.

In the case of TFIIIA a poly finger system is build up. The distance between fingers is more variable and DNA is apparently bent.

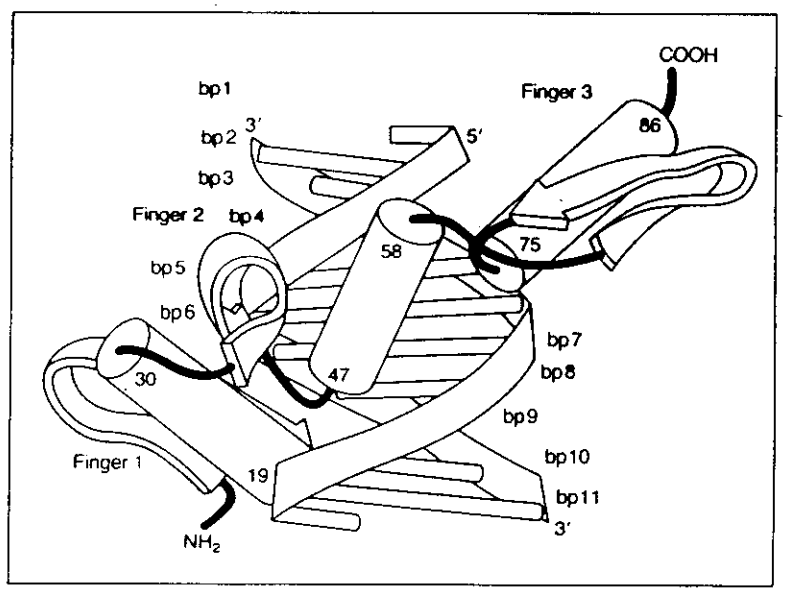
FIG. 12

ZINC FINGERS

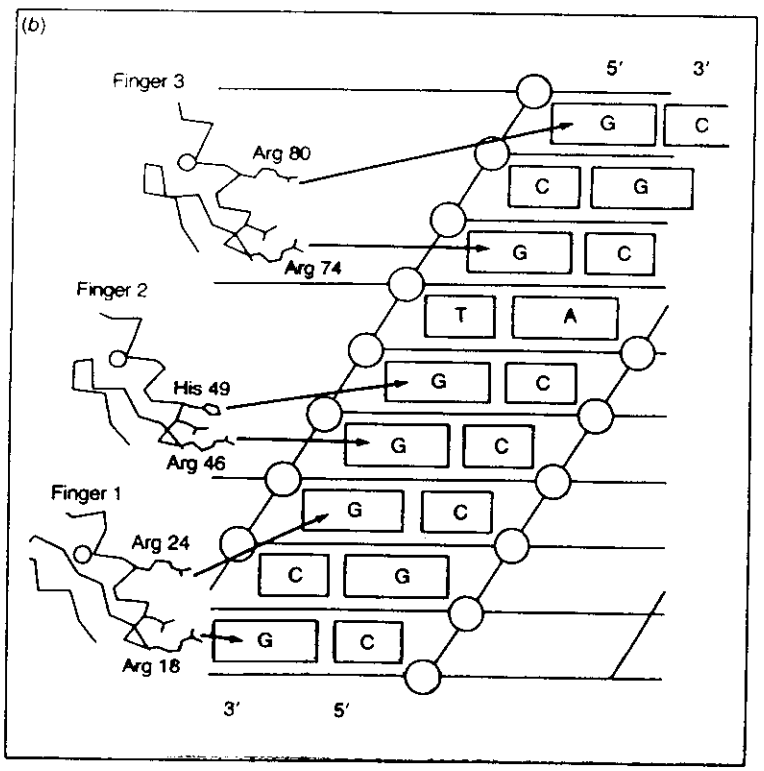


β1 β2 α1

● CONTACTS WITH BACKBONE
● CONTACTS IN THE MAJOR GROOVE
} ZIF 268



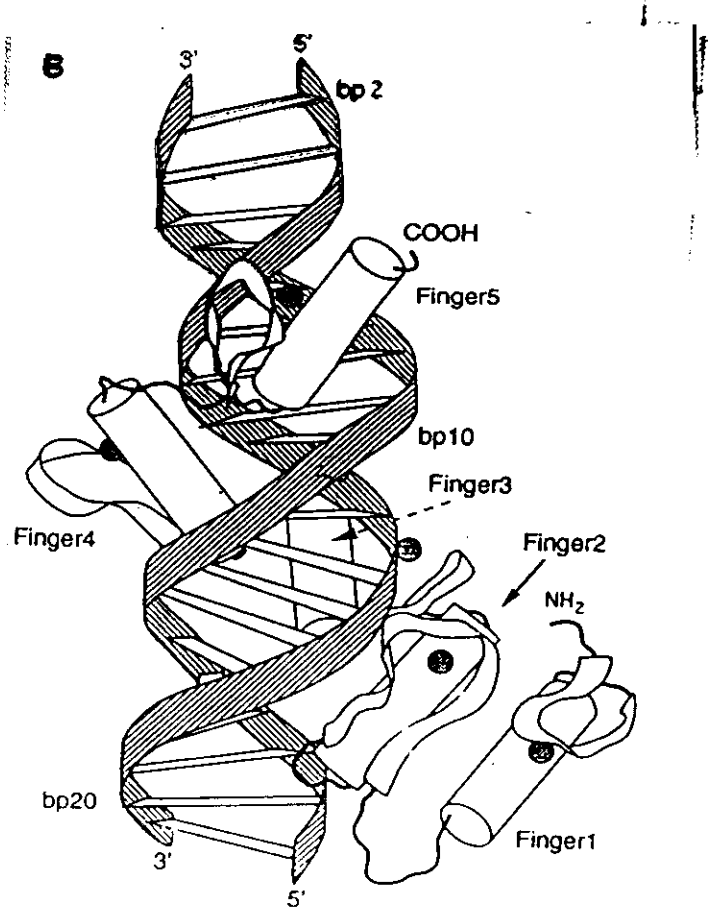
ZIF/DNA



ZIF 268/DNA

FIG. 13

GLI / DNA COMPLEX



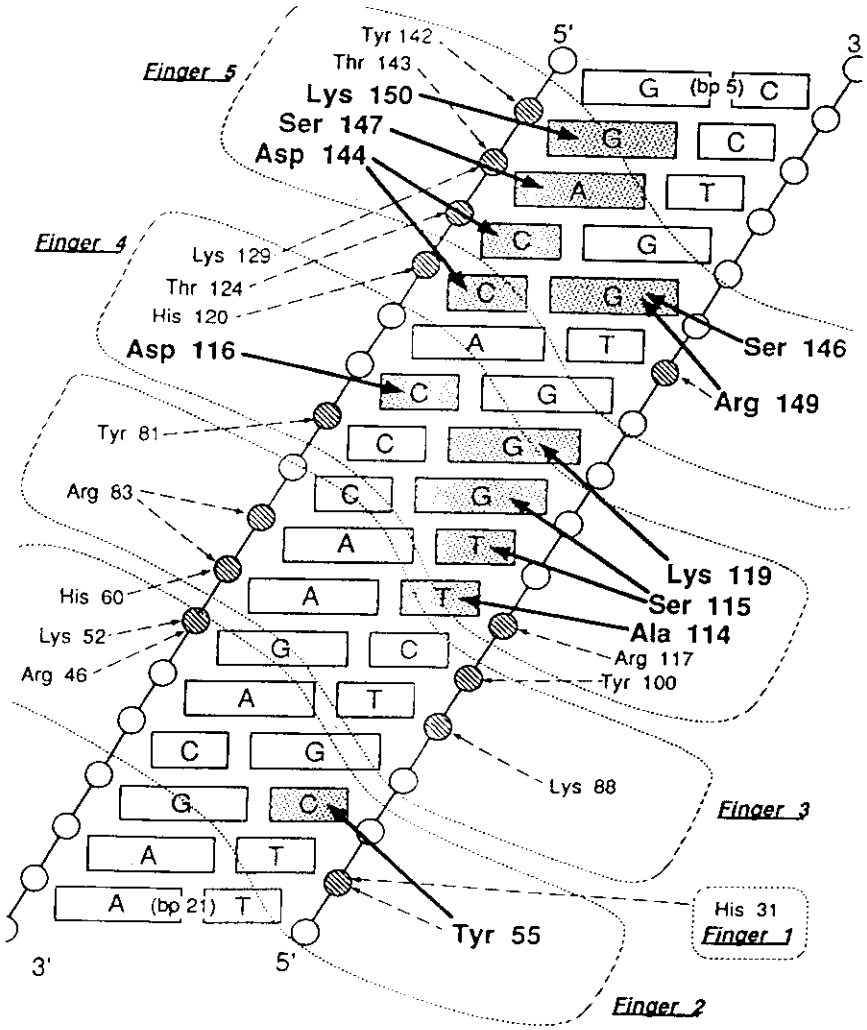


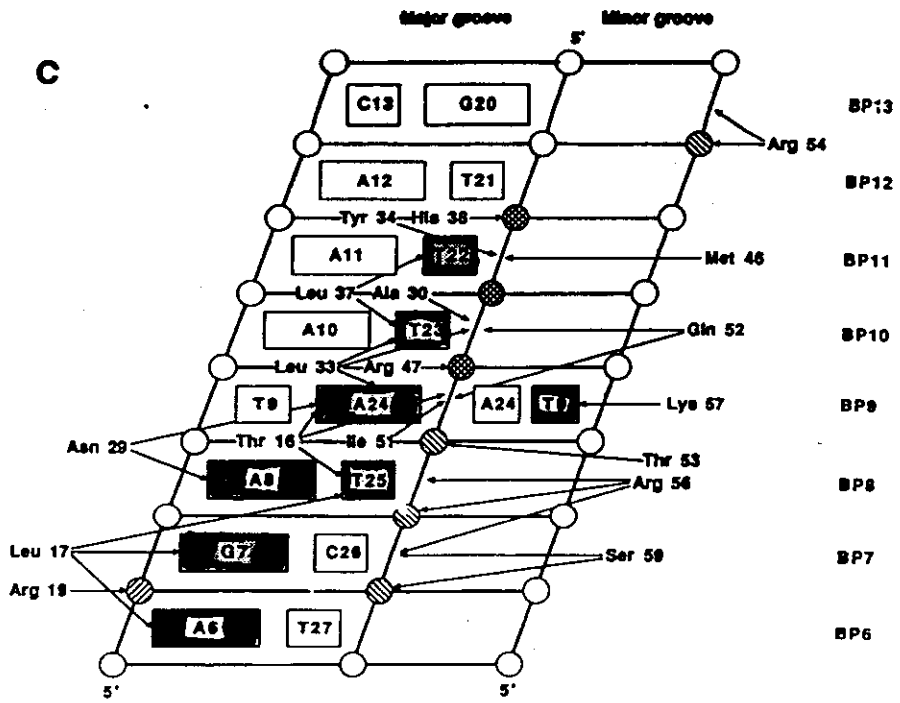
FIG. 14

GATA-1

The zinc binding core consists of two short, irregular two-stranded β-sheets connected by a loop and followed by an α-helix. The structure is somewhat similar to the N-terminal zinc binding module of glucocorticoid receptor (GR). However the two α-helices contact the major groove at different angles.
GATA-1 binds DNA at with high affinity using a single metal core. Nuclear hormone receptors and GAL-4 bind as dimers, while C_2H_2 class uses tandem arrays. This is probably due to the C-terminal DNA-contacting arm.

Nuclear hormone receptor DBD

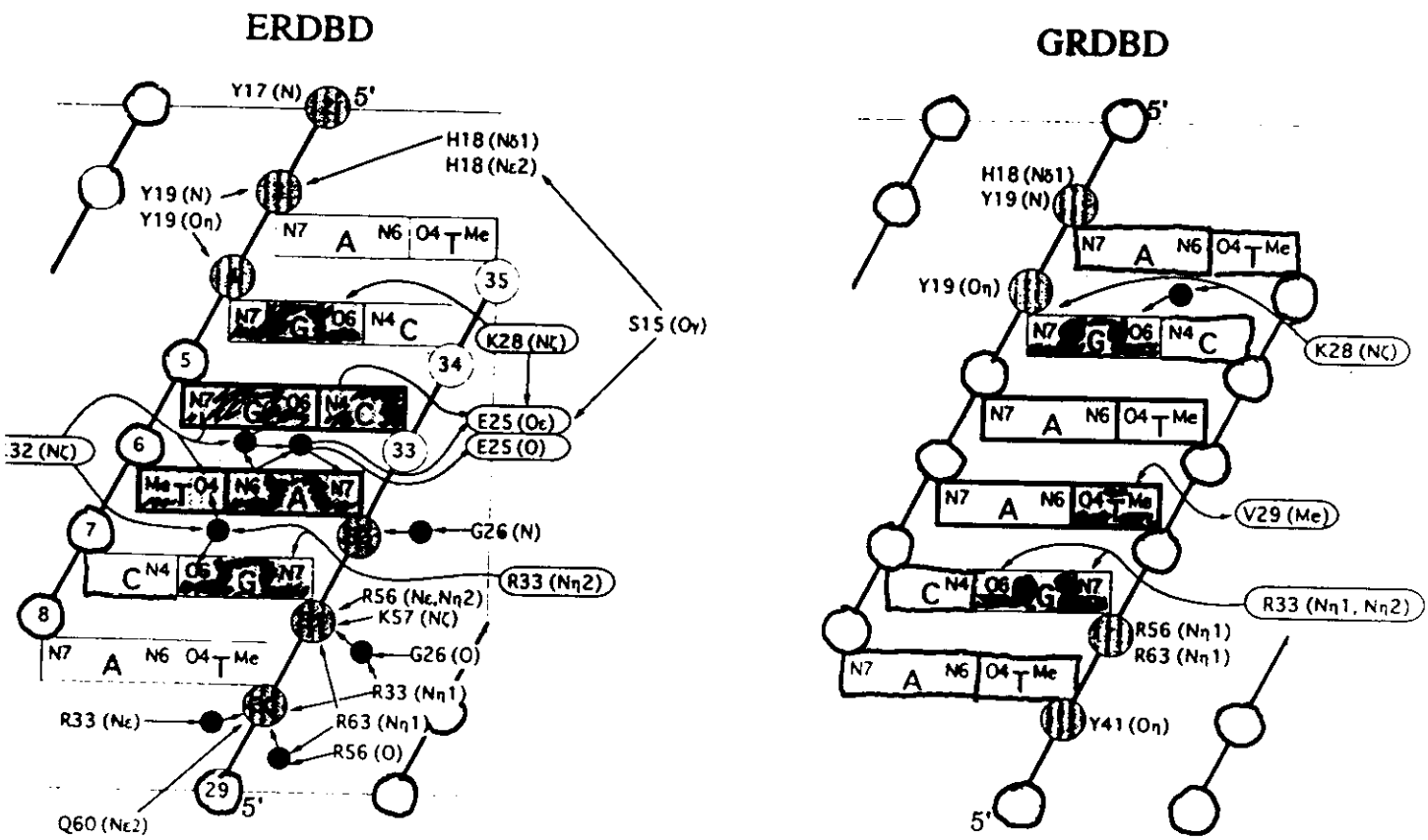
Steroid receptors are an important family of regulatory proteins. They contain separate domains for hormone binding, DNA binding and for transcriptional activation. The DBD, which contains about 70 residues, have 8 conserved cysteine residues. Unlike other different zinc motifs the 8 cysteine do not form two separate fingers but a unique, separate structural unit. This consists of a single globular domain with a pair of α-helices. The two helices are roughly perpendicular and held together by hydrophobic contacts. The two zinc atoms bind near the start of each helix and holds a peptide loop against the C-terminal end of the helix.
The crystal structure of the glucocorticoid/DNA complex shows that a peptide binds as a dimer. The first helix of each subunit fits into the major groove. The second helix makes backbone contacts and provides the dimerization interface.
A chosen receptor DBD behaves similarly to GR-DBD. FIG. 15



GATA DNA BINDING DOMAIN / DNA COMPLEX

BASES IN RED ARE CONTACTED BY AMINOACIDS
ALL THE CONTACTS IN THE MAJOR GROOVE INVOLVE THE HELIX AND THE LOOP CONNECTING β2 AND β3 STRANDS - THE MAJORITY OF THESE BASE CONTACTS ARE HYDROPHOBIC

FIG.16



BASES IN RED ARE CONTACTED

● WATER MOLECULES

DNA INDUCES DIMERIZATION OF PROTEIN AT LEAST FOR ER

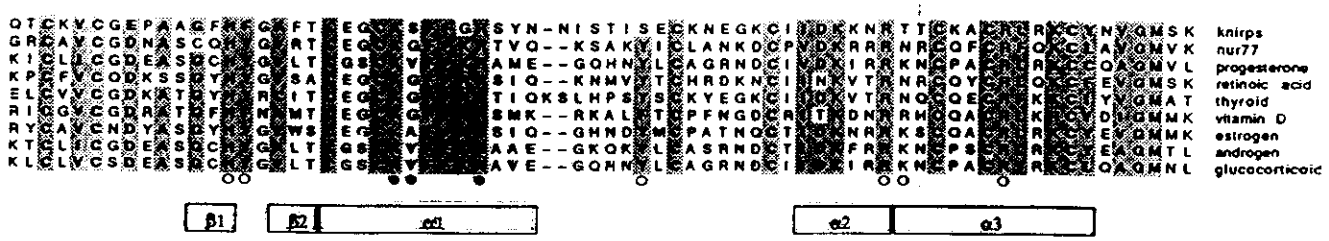


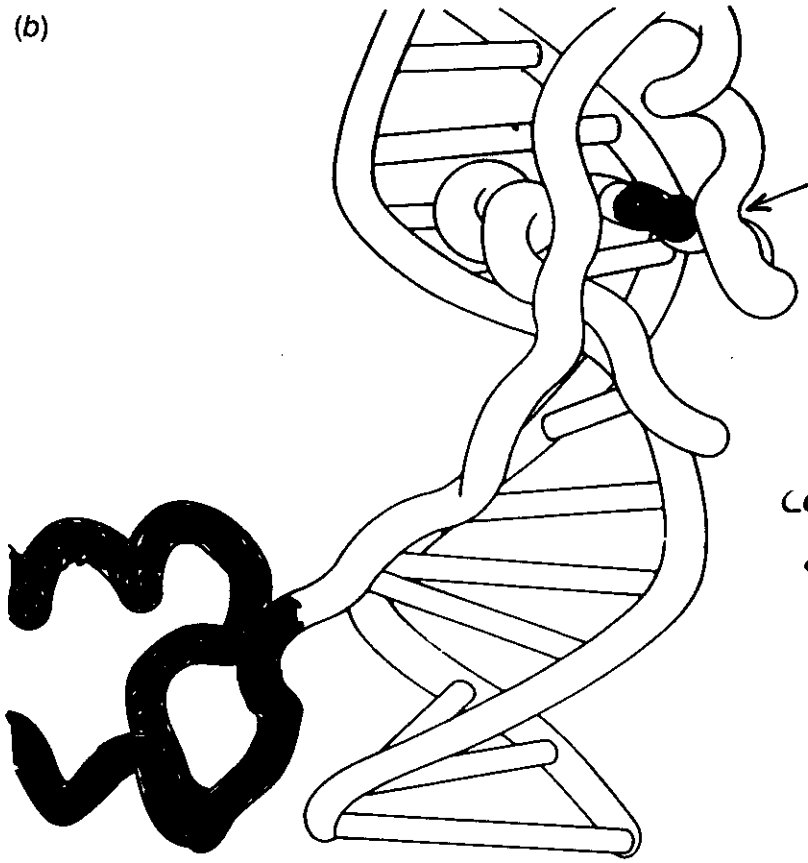
FIG. 17

GAL4-C_6

This protein has a DBD consisting of a Zn binuclear cluster. The contact with DNA is established by amino acids contained within a short α-helical region in the major groove. Many contacts are however made with the backbone with the amino acids in the more extended chain. The GAL4 protein, like the steroid hormone receptors, binds to its recognition site as a dimer with the two monomeric units being held together by a short-coiled coil of two parallel α-helices.

FIG. 18

(b)


ZINC ATOMS
CGGAGGACTGTCCTCCG
CONSENSUS SEQUENCE

(c)

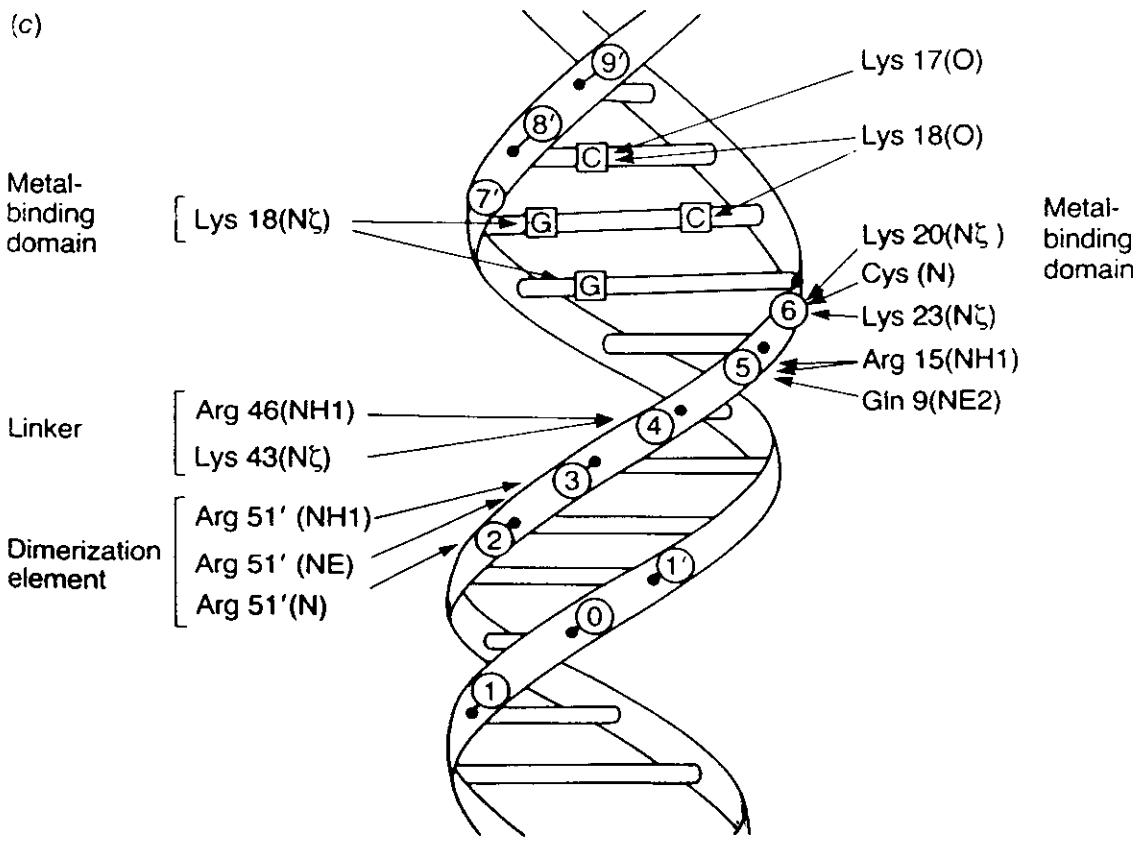
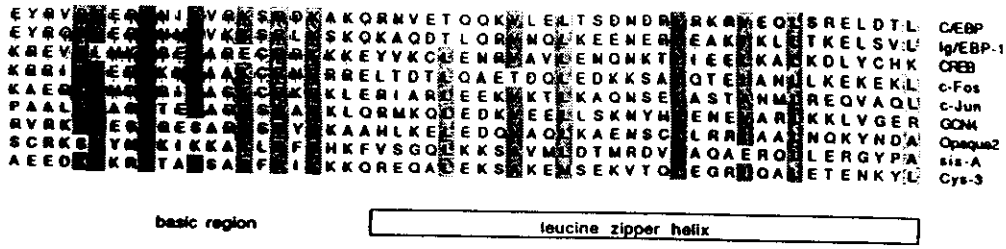
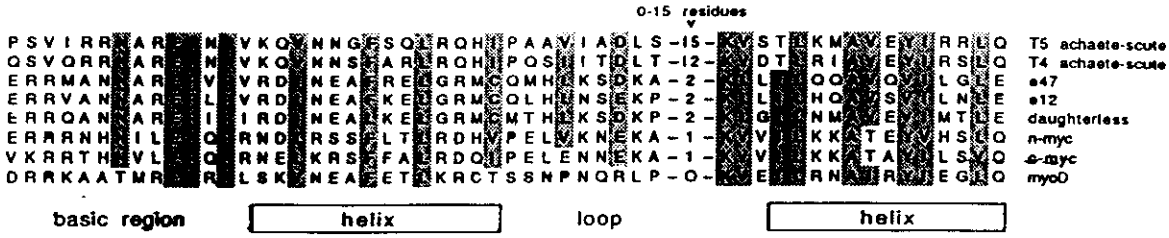

Metal-binding domain
Lys 18(Nζ)
Linker
Arg 46(NH1)
Lys 43(Nζ)
Dimerization element
Arg 51′ (NH1)
Arg 51′ (NE)
Arg 51′(N)
Lys 17(O)
Lys 18(O)
Lys 20(Nζ)
Cys (N)
Lys 23(Nζ)
Arg 15(NH1)
Gln 9(NE2)
Metal-binding domain
9′
8′
7′
6
5
4
3
2
1′
0
1
C
G
C
G

FIG. 19


C/EBP
Ig/EBP-1
CREB
c-Fos
c-Jun
GCN4
Opaque2
sis-A
Cys-3
basic region
leucine zipper helix


0-15 residues
T5 achaete-scute
T4 achaete-scute
e47
e12
daughterless
n-myc
c-myc
myoD
basic region
helix
loop
helix

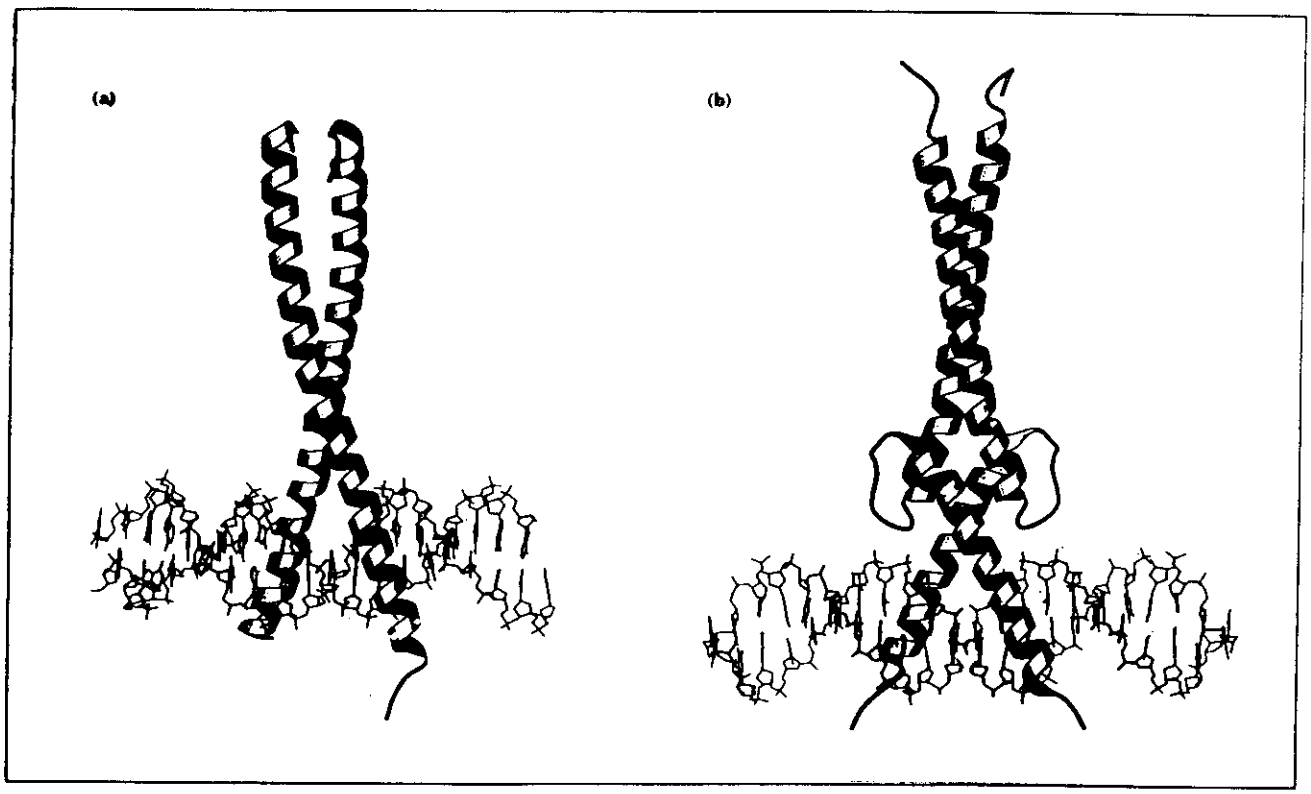

(a)
(b)

FIG. 20

BZIP AND BLHL DOMAINS

The basic-leucine zipper and basic helix-loop-helix proteins have important roles in differentiation and development. They also are interesting because they illustrate the important roles that heterodimer formation can play in the regulation of gene expression.

The basic region (the amino terminal region) consists of two extended α-helices rich in basic residues. The basic region is unstructured in solution and α-helical when bound to DNA. The residues in the inner face of the α-helices contact the base-pairs in the major groove, whereas the neighbouring residue contact the phosphodiester backbone.

LZ and HLH dimerization segments contribute to DNA binding specificity by determining which subunits form stable dimers, and by appropriately positioning the basic region helices over the binding sites.

Due to the lack of constraints of globular domains the basic region is more flexible and more adaptable than other recognition helices.

The leucine zipper is responsible of the dimerization. Apolar residues located in the inner face together with polar residue make homodimerization or heterodimerization (c-jun and c-fos) more stable

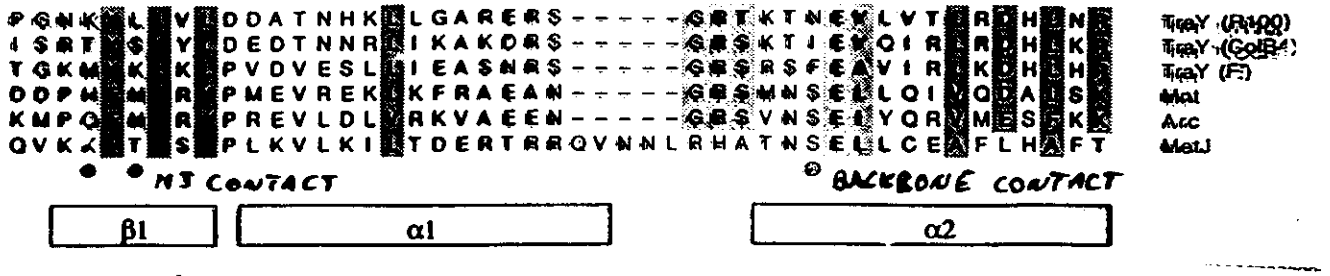
The bHLH dimer is a left-handed, parallel four helix bundle. It adopts the same strategy as the LZ dimer. Some bHLH contain a LZ for efficient dimerization

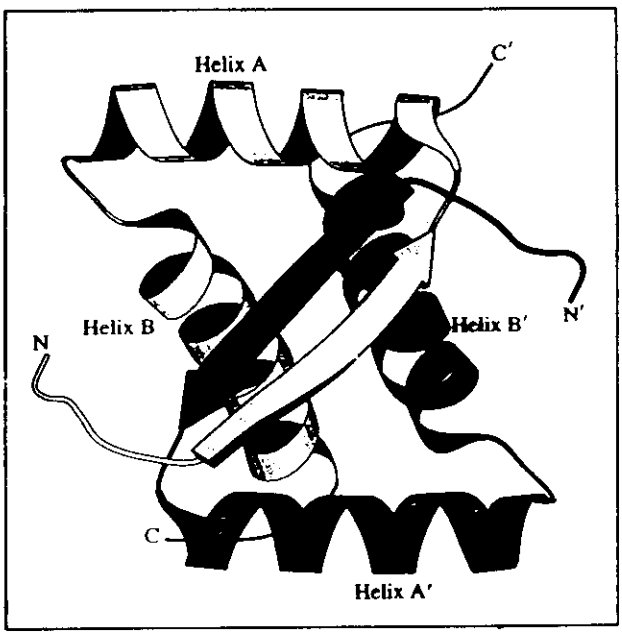
FIG. 21

The variability of the recognition by the basic region is demonstrated by the crystal structures of GCN4 with two sites: CRE site and AP-1 site, which differ in the spacing of the two half-sites. In the case of the CRE site there is a slight increase of the divergence of the two α-helices paralleled by a bent (20°) of the DNA towards the LZ.

Most bLHL proteins bind the DNA hexamer CANNTG where N are C or G.

FIG. 22





RIBBON-HELIX-HELIX

DIMER OF ARC βαα motif

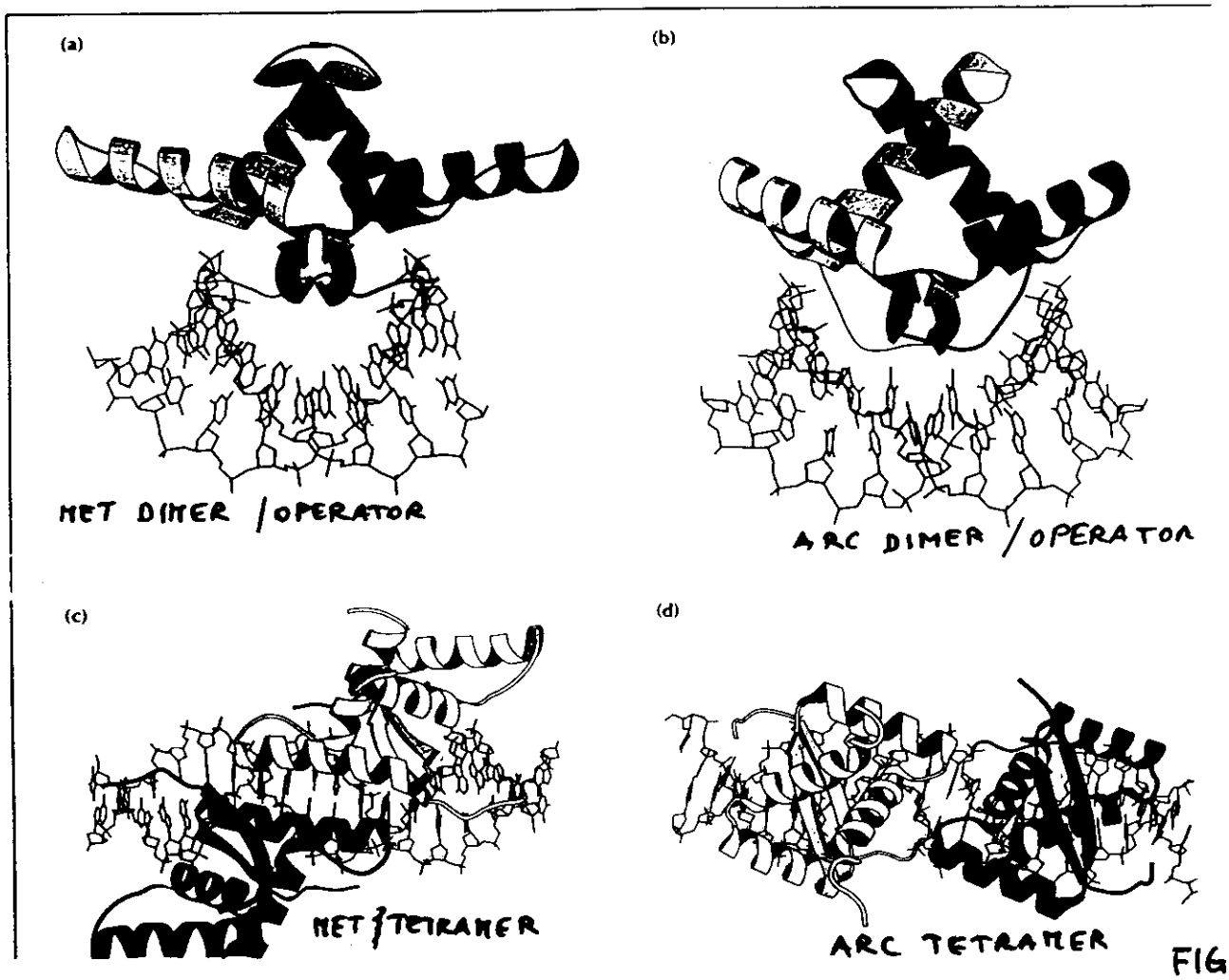
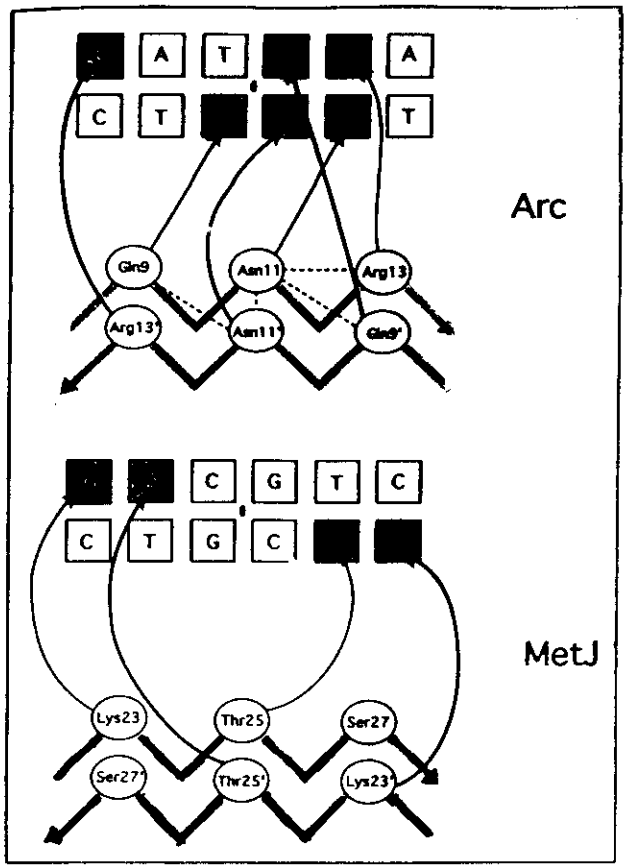
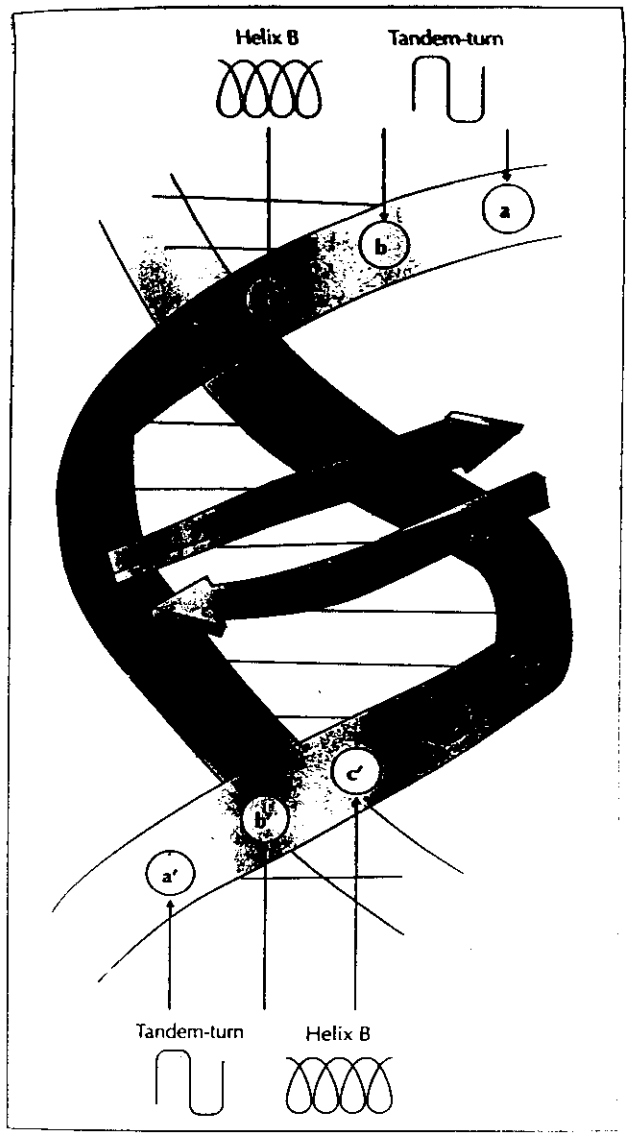


FIG. 2

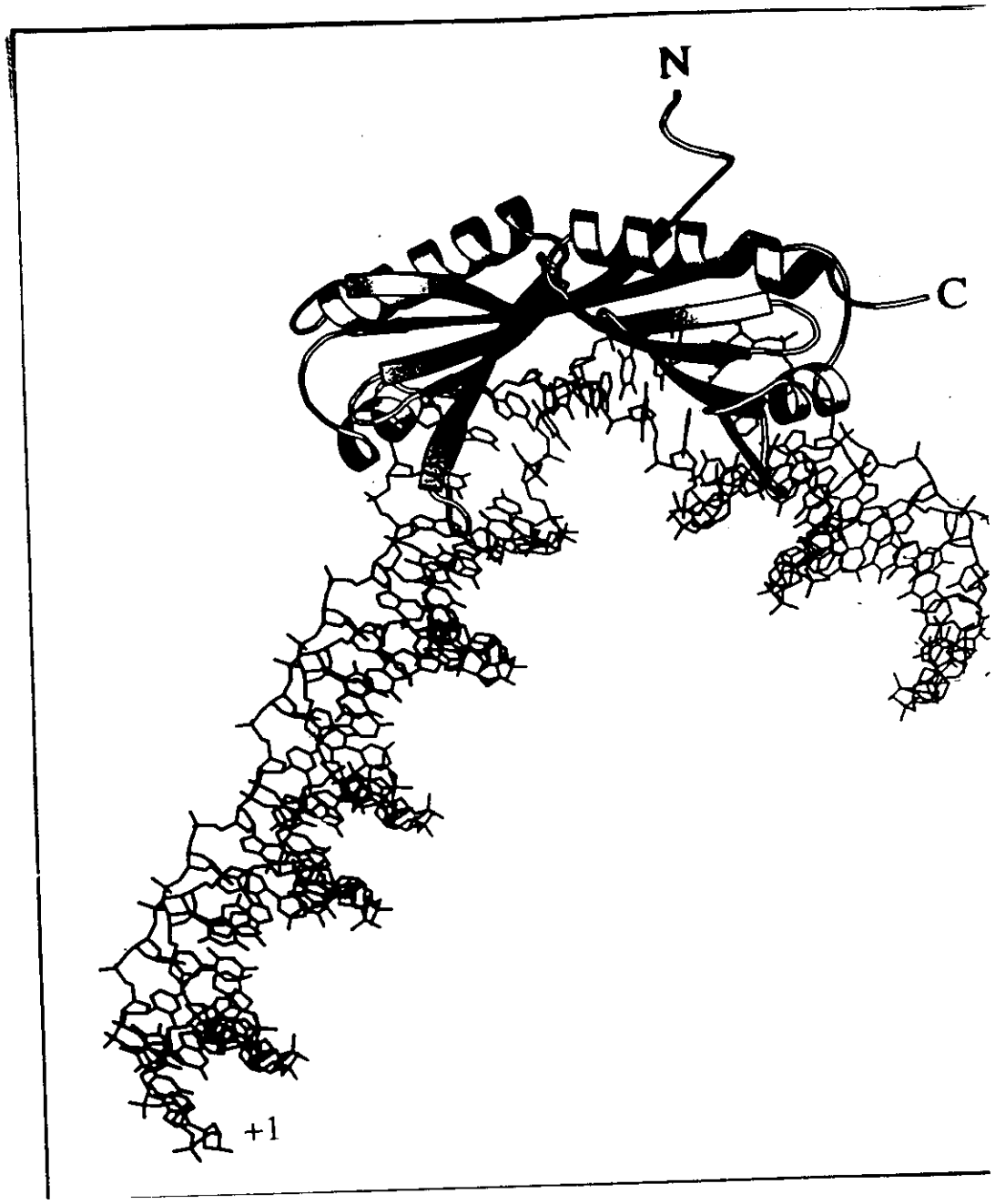
RIBBON - HELIX - HELIX





CONTACTED ONLY BY MET J
CONTACTED ONLY BY ARC

FIG. 24



TBP2 FROM ARABIDOPSIS THALIANA COMPLEXED WITH ADENOVIRUS MAJOR LATE PROMOTER TATA ELEMENT

KIM J.L. ET AL. NATURE 365, 520-527 (1993)

THE C-TERMINAL PORTION CONTAINS 180 aa WITH TWO STRUCTURAL REPEATS

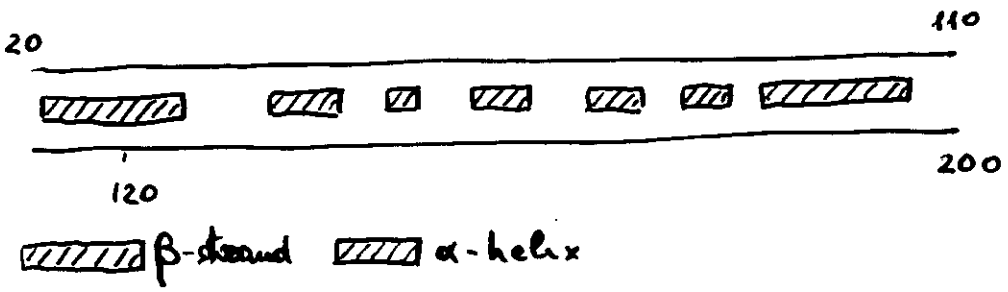
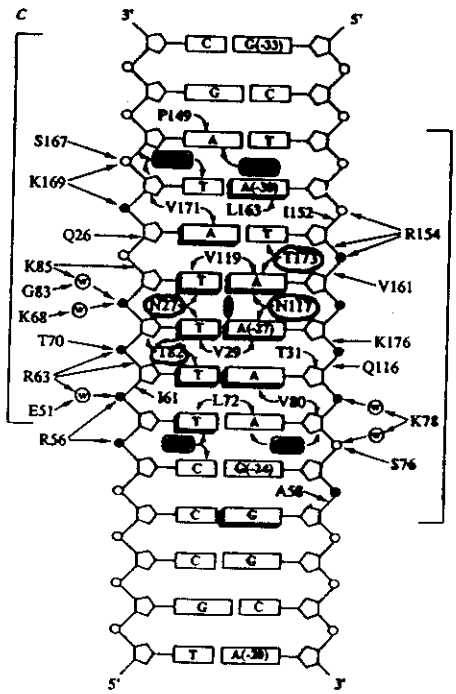
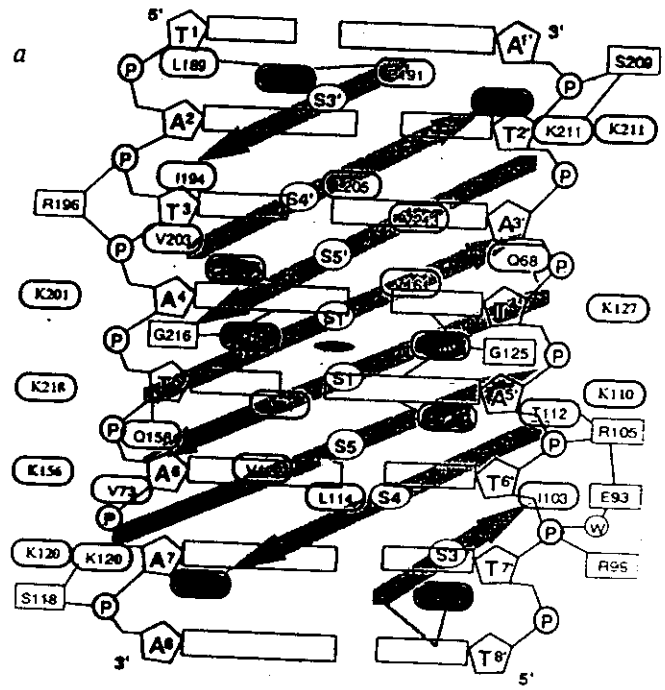


FIG. 25



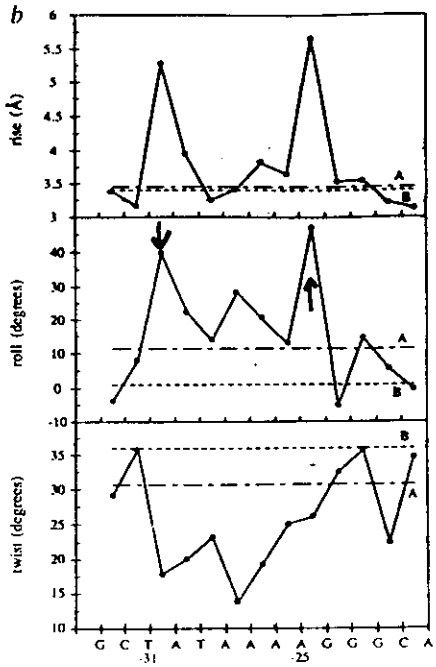
TBP2 / TATA BOX



yTBP / TATA box

The interface is primarily hydrophobic - only 1080 Å^2 over 3142 Å^2 are hydrophilic (six hydrogen bonds)
Phe residues responsible of the kinks are in red
Hydrogen bonds to base pairs are made by aminoacids shown in green

FIG.26



TBP2 / TATA box

Base number	0	1	2	3	4	5	6	7	8	9	10	11
Helical twist (°)	32.85	30.90	18.12	22.71	3.05	20.72	20.86	22.85	37.34	31.93	43.50	
Roll (°)	2.12	45.30	19.57	13.82	25.11	18.78	15.24	37.59	2.09	1.59	4.57	
Slide (Å)	-1.61	-0.80	-0.91	1.63	0.79	1.67	0.40	0.51	-0.11	-1.23	0.07	
Rise (Å)	3.47	5.49	3.71	3.22	3.78	3.59	3.74	5.32	3.35	3.50	3.65	
P-P distance (Å)		5.63	6.27	5.50	5.75	5.80	6.24	5.75	6.70	7.04	6.58	7.04
Minor groove width (Å)	5' G	T	A	T	A	T	A	A	A	A	C	G
	5.17	7.32	8.07	9.72	9.79	9.73	8.13	7.91	5.84	4.78	5.44	
	3' C	A	T	A	T	A	T	T	T	T	G	C
P-P distance (Å)		6.82	5.64	6.09	5.74	5.92	5.87	6.40	6.02	6.60	6.62	6.67
Buckle (°)	-10.10	-14.04	-38.77	-28.26	-5.55	11.03	23.02	24.49	6.67	12.35	5.20	-3.35
Propeller twist (°)	-16.37	-13.27	5.05	-22.22	-10.32	-7.41	-24.83	-37.71	-0.37	-13.39	-15.80	-5.98

yTBP / TATA box

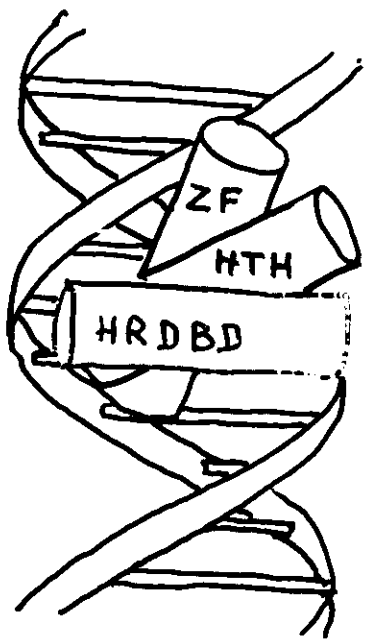
There is a marked reduction of twist angles compensated by a positive writhe.
The unwinding and the large positive roll angles considerably broaden the minor groove from 7.7 Å to 10 Å.
Major groove is deepened and an angle of about 100° is created
Minor groove is open and convex whereas the protein is concave like a saddle.

FIG.27

PRINCIPLES OF RECOGNITION

1. MOST PROTEINS USE α-HELICES AS RECOGNITION STRUCTURE. HOWEVER IT IS DANGEROUS TO THINK THAT A SINGLE HELIX WOULD BE ABLE TO BIND SELECTIVELY TO THE DNA SITE. IN ALL THE COMPLEX SO FAR ESAMINATED THE BINDING SPECIFICITY RESULTS FROM A SET OF INTERACTIONS.

   MOREOUER THE ORIENTATION OF THE HELIX WITH RESPECT TO THE MATOR GROOVE IS VARIABLE IN THE DIFFERENT CLASSES OF PROTEINS



THE ORIENTATION OF THE HELIX IS VERY MUCH DETERMINED BY THE SURROUNDING REGIONS OF THE PROTEIN

FIG. 28

2. THE INTERACTIONS WITH THE BASES PLAY A CRUCIAL ROLE IN SITE SPECIFIC BINDING. THE CONTACTS ARE

a) DIRECT HYDROGEN BONDS BETWEEN SIDE CHAINS AND BASES
b) OCCASIONAL HYDROGEN BONDS POLYPEPTIDE BACKBONE AND BASES
c) HYDROGEN BONDS MEDIATED BY WATER MOLECULES
d) HYDROPHOBIC CONTACTS

MAJOR GROOVE IS BY FAR THE MOST USED FOR THESE CONTACTS, ALSO IF SIGNIFICANT EXCEPTIONS ARE KNOWN THIS IS DUE BOTH TO THE HIGHER ACCESSIBILITY OF MAJOR GROOVE AND TO THE HIGHER VARIABILITY OF THE CONTACTABLE MOIETIES IN THE BASE WEDGES

THERE IS NO RECOGNITION CODE, i.e. NO CORRESPONDENCE BETWEEN A BASE AND AN AMINO-ACID SIDE-CHAIN. PURINE ARE MAJOR TARGETS (ESPECIALLY GUANINE)
HOWEVER IN SOME CLASS OF PROTEIN CONSERVED AMINOACIDS MAKE CONTACTS WITH THE SAME BASE
THIS SUGGESTS THAT A GIVEN FAMILY OF PROTEINS MAY HAVE A CONSERVED DOCKING MECHANISM AND A CONSERVED SET OF CONTACTS

3. CONTACTS WITH DNA BACKBONE ARE VERY IMPORTANT AND OFTEN ARE MORE NUMEROUS THAN DIRECT BASE CONTACTS THEY MAY HELP TO HOLD THE PROTEIN IN A FIXED ARRANGEMENT AGAINST THE BASES, INDIRECTLY ENHANCING THE INTERACTION SPECIFICITY

FIG. 29

4. BINDING SPECIFICITY APPEARS OFTEN RELATED TO DNA STRUCTURE AND/OR FLEXIBILITY

a) A GIVEN BINDING SITE MAY HAVE A STRUCTURE WHICH DEVIATES FROM A CANONICAL B-DNA AND WHICH IS A PREREQUISITE FOR PROTEIN BINDING

b) A GIVEN BINDING SITE HAS A SEQUENCE WHICH CONTRIBUTES TO SPECIFIC BINDING WITH ITS FLEXIBILITY

FIG. 30