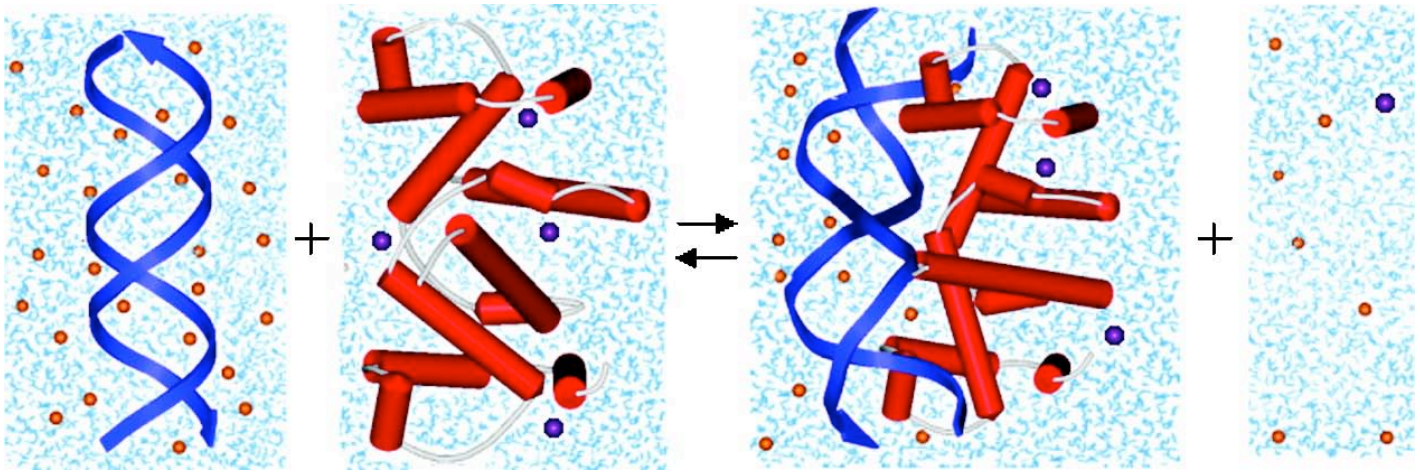
Protein-nucleic acid interactions

October 6, 2009

Professor Wilma K. Olson

Schematic representation of protein-DNA binding

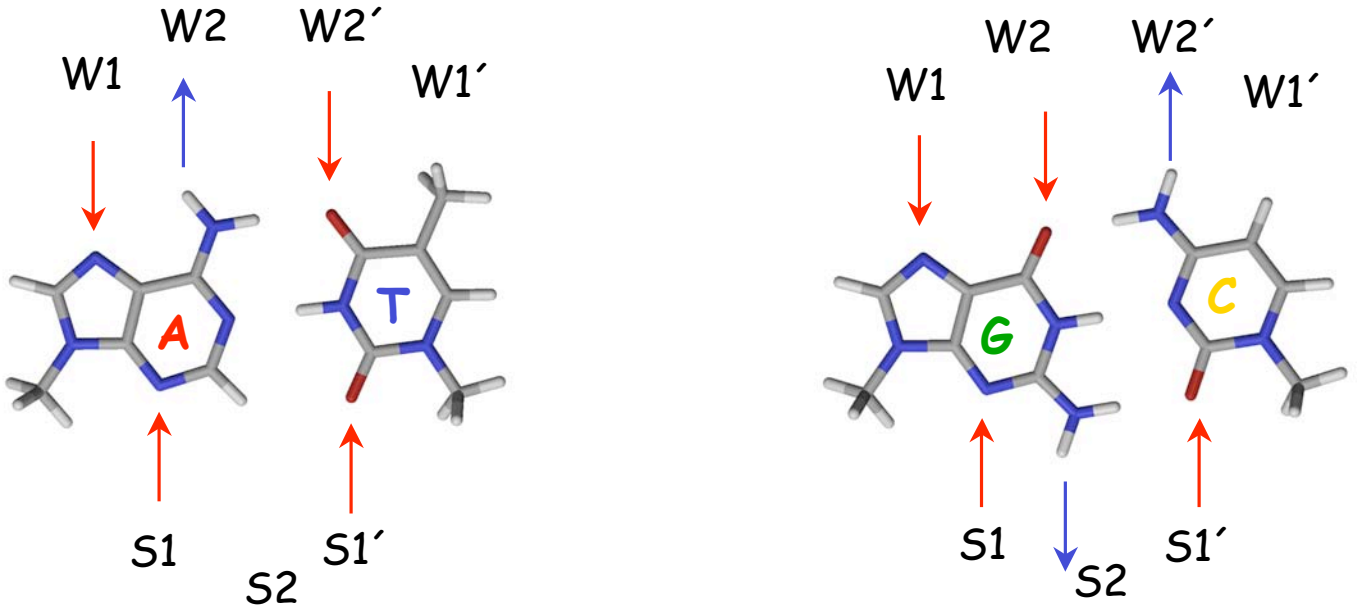


The binding process is accompanied by the release of water molecules (blue) and counterions (red) as well as changes in DNA and protein conformations. Waters remaining at the interface facilitate binding by screening electrostatic repulsions between like charges of the protein and the DNA. A small fraction of the interfacial waters form extended hydrogen bonds between the protein and the DNA, thereby compensating for the lack of direct hydrogen bonds.

Jayaram & Jain (2004) "The role of water in in protein-DNA recognition." *Ann. Rev. Biophys. Biomol. Struct.* 33, 343-361

Nucleic-acid recognition principles

The H-bond donor and acceptor atoms on the major-groove edges of the Watson-Crick base pairs suggest a mechanism of direct sequence recognition.



W - potential recognition site in **wide** major groove

S - potential recognition site in **small** minor groove

Nitrogen atom

Oxygen atom

Seeman *et al.* (1976) "Sequence-specific recognition of double helical nucleic acids by proteins." *Proc. Natl. Acad. Sci. USA* 73, 804-808

The major-groove H-bond donor and acceptor patterns present unique motifs for direct recognition of the four Watson-Crick pairs.

Base pair	Major groove				Minor groove		
	W1	W2	W2'	W1'	S1	S2	S1'
A·T	N	NH2	O	C-Me	N	H	O
T·A	C-Me	O	NH2	N	O	H	N
G·C	N	O	NH2	C-H	N	NH2	O
C·G	C-H	NH2	O	N	O	NH2	N

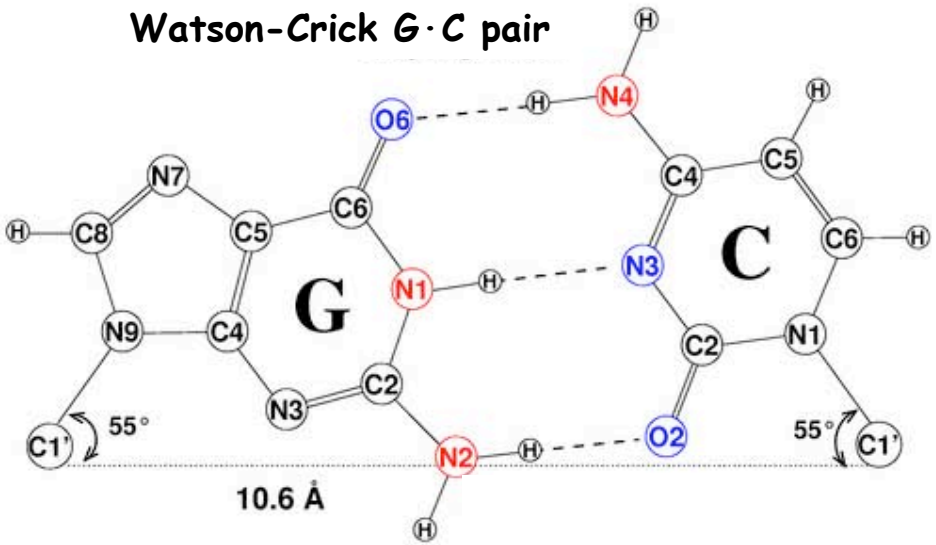
H-bond donor H-bond acceptor

Bidentate hydrogen bonds are sufficient to discriminate the base pairs.

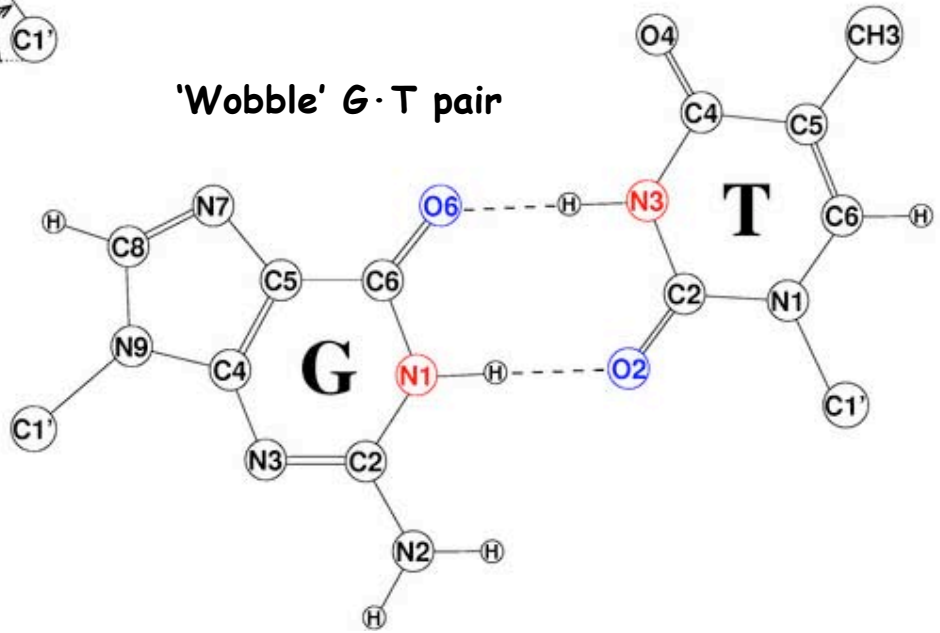
Seeman *et al.* (1976) "Sequence-specific recognition of double helical nucleic acids by proteins." *Proc. Natl. Acad. Sci. USA* 73, 804-808

The minor-groove H-bond donor and acceptor patterns discriminate Watson-Crick from non-canonical base pairs.

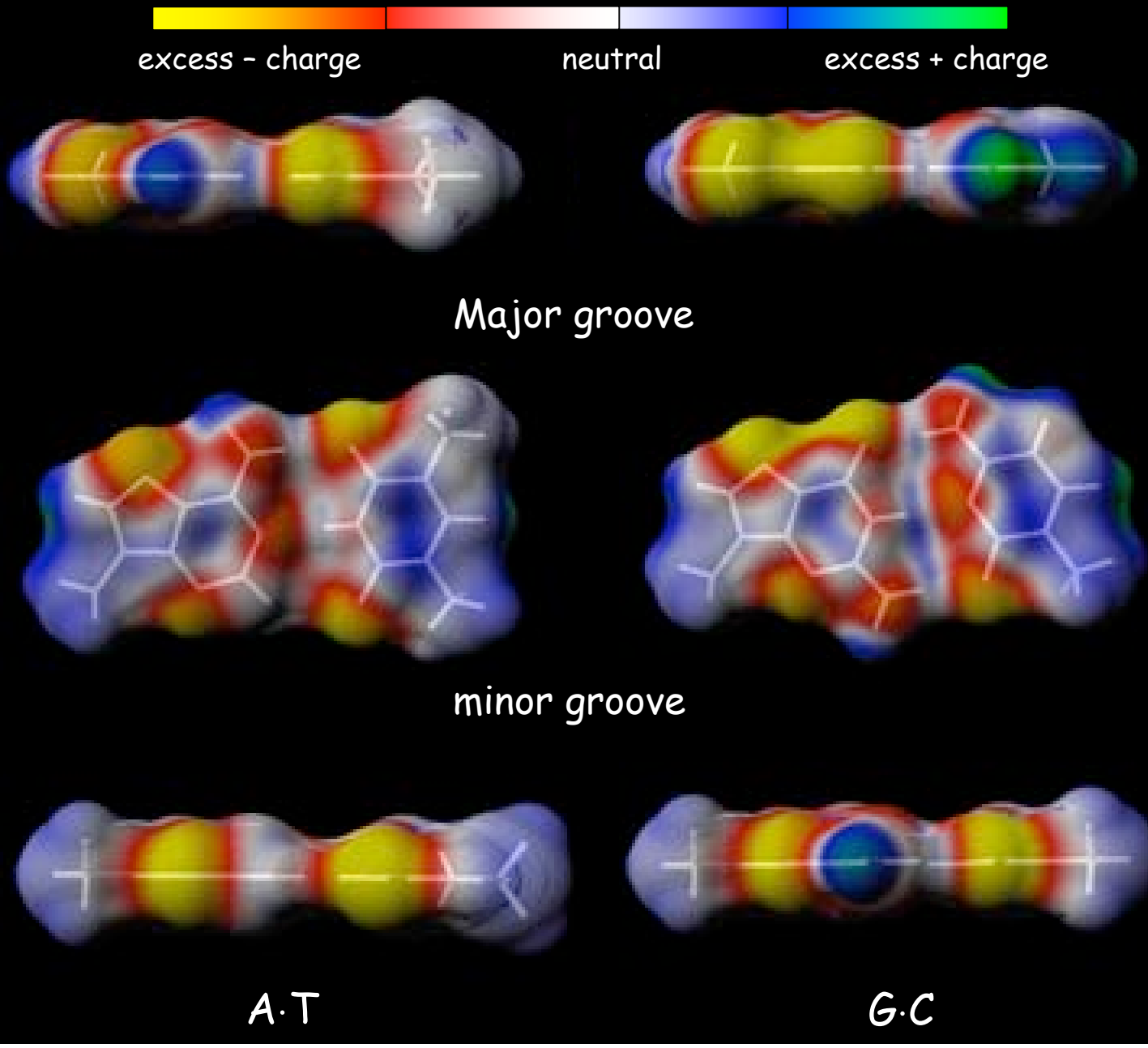
Watson-Crick G·C pair



'Wobble' G·T pair



The distribution of electronic charge underlies the base-pair recognition pattern.



DNA-protein contact propensities

The distribution of H bonds between DNA and amino-acid atoms in protein-DNA crystal complexes depends upon sequence.

Observed (expected) numbers in 129 non-redundant structures

Amino acids			DNA bases								DNA backbone				Total	
			Thymine		Cytosine		Adenine		Guanine		Sugar		Phosphate			
Arginine	ARG	R	**24**	**(2.5)**	8	(2.0)	**19**	**(4.2)**	**98**	**(4.0)**	8	(1.9)	**218**	**(49.9)**	**375**	**(64.7)**
Lysine	LYS	K	**9**	**(4.4)**	6	(3.4)	**4**	**(7.4)**	**30**	**(7.1)**	3	(3.2)	**109**	**(86.7)**	**165**	**(112.6)**
Serine	SER	S	3	(3.0)	2	(2.2)	1	(5.0)	12	(4.6)	2	(2.1)	**91**	**(57.3)**	**113**	**(74.4)**
Threonine	THR	T	5	(2.4)	3	(2.0)	4	(4.2)	-	(4.0)	1	(1.8)	**79**	**(49.2)**	**92**	**(63.9)**
Asparagine	ASN	N	7	(3.6)	10	(2.7)	**18**	**(6.0)**	7	(5.8)	3	(2.6)	**43**	**(70.7)**	88	(91.8)
Glutamine	GLN	Q	2	(2.2)	2	(1.7)	**16**	**(3.8)**	6	(3.6)	2	(1.7)	**42**	**(44.8)**	70	(58.1)
Glycine	GLY	G	1	(3.2)	4	(2.4)	-	(5.4)	6	(5.1)	1	(2.4)	29	(63.3)	**41**	**(82.2)**
Histidine	HIS	H	-	(0.8)	1	(0.6)	1	(1.5)	12	(1.4)	-	(0.7)	26	(18.3)	**40**	**(23.7)**
Tyrosine	TYR	Y	-	(1.2)	2	(1.0)	-	(2.1)	1	(2.0)	1	(1.0)	35	(25.7)	39	(33.4)
Alanine	ALA	A	1	(2.5)	1	(2.0)	-	(4.2)	1	(4.0)	-	(1.9)	**17**	**(49.8)**	**20**	**(64.6)**
Glutamate	GLU	E	-	(3.8)	**10**	**(3.0)**	1	(6.5)	1	(6.2)	-	(2.9)	**6**	**(76.2)**	**18**	**(99.0)**
Isoleucine	ILE	I	-	(0.7)	-	(0.5)	-	(1.3)	3	(1.2)	-	(0.6)	11	(15.9)	14	(20.7)
Aspartate	ASP	D	-	(4.5)	**5**	**(3.4)**	2	(7.5)	2	(7.2)	-	(3.3)	**2**	**(88.3)**	**11**	**(114.7)**
Valine	VAL	V	-	(1.2)	-	(1.0)	-	(2.0)	-	(2.0)	-	(0.9)	**8**	**(24.5)**	**8**	**(31.8)**
Cysteine	CYS	C	-	(0.2)	1	(0.2)	-	(0.5)	-	(0.5)	-	(0.3)	4	(6.7)	5	(8.7)
Phenylalanine	PHE	F	-	(0.6)	-	(0.5)	-	(1.1)	1	(1.1)	-	(0.5)	4	(14.4)	5	(18.6)
Leucine	LEU	L	-	(1.5)	-	(1.1)	-	(2.6)	-	(2.5)	-	(1.2)	**5**	**(30.8)**	**5**	**(40.0)**
Methionine	MET	M	1	(0.4)	-	(0.3)	-	(0.7)	-	(0.7)	-	(0.3)	3	(9.1)	4	(11.8)
Tryptophan	TRP	W	-	(0.3)	-	(0.2)	-	(0.7)	-	(0.6)	-	(0.3)	3	(8.7)	3	(11.3)
Proline	PRO	P	-	(3.5)	1	(2.7)	-	(6.0)	-	(5.7)	-	(2.6)	**-**	**(70.0)**	**1**	**(90.9)**
Total			53	(42.5)	**56**	**(33.0)**	66	(73.4)	**180**	**(69.4)**	21	(32.2)	**735**	**(860.3)**	1,111	(1,111)

Luscombe *et al.* (2001) "Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level." *Nucleic Acids Res.* 29, 2860-2874

Proteins 'read' DNA primarily through the phosphate and base atoms using amino acid residues which are positively charged or polar.

		Protein residues			
DNA	Sample size	Arg, Lys +	Asp, Glu –	Ala, Ile, Leu, Met, Phe, Pro, Val	Asn, Cys, Gln, His, Ser, Thr, Trp, Tyr
Base	4266	**1061**	115	108	**788**
Phosphate	3853	**1445**	61	166	**1301**
Sugar	4263	345	41	80	303

Number of close (≤ 3.4 Å) contacts between protein and DNA atoms in 132 unique protein-DNA complexes

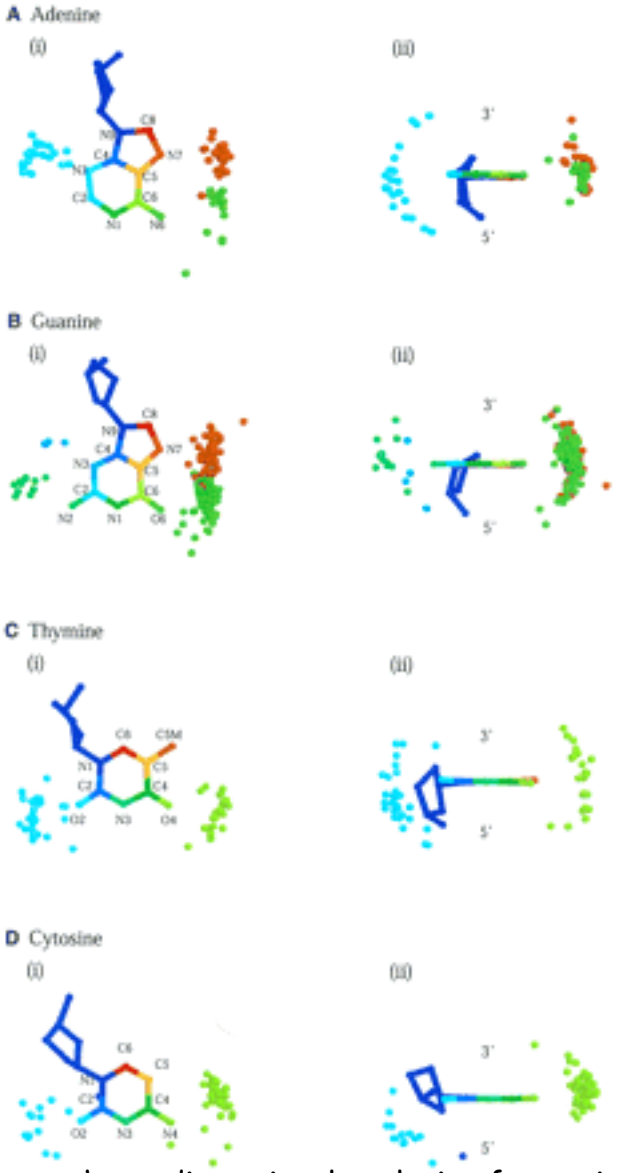
Yun Li

Spatial interaction patterns

H-bonded amino-acid atoms localize in tight clusters around the DNA bases.

Interacting amino-acid atoms superposed on *A*, *G*, T, *C* and identified by the same colors: major groove - red/green; minor groove - cyan/green.

Distributions shown in two orientations: (left) along the base normal from the 3'-end and (right) toward the base-pairing edge.

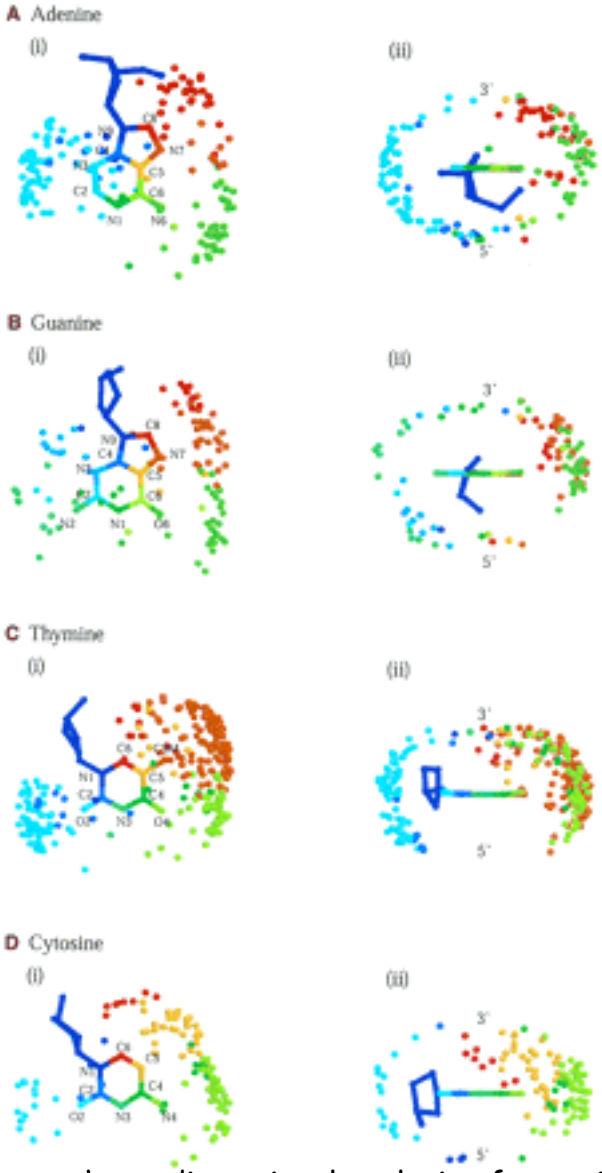


Luscombe *et al.* (2001) "Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level." *Nucleic Acids Res.* 29, 2860-2874

The van der Waals' contacts of amino acid and base are more distant and dispersed.

Interacting amino-acid atoms superposed on *A*, *G*, T, *C* and identified by the same color: major groove - red/green; minor groove - cyan/green.

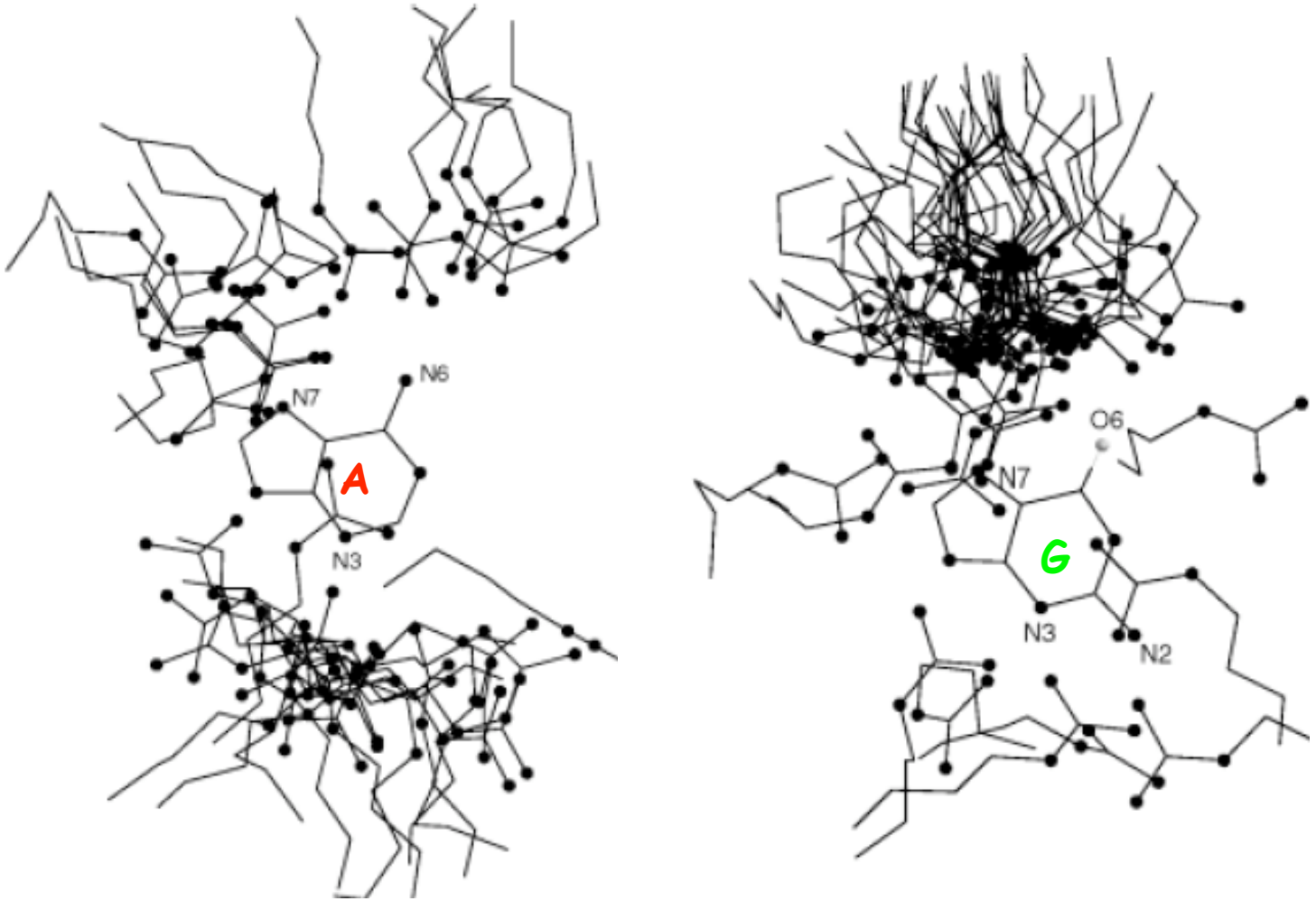
Distributions shown in two orientations: (left) along the base normal from the 3'-end and (right) toward the base-pairing edge.



Luscombe *et al.* (2001) "Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level." *Nucleic Acids Res.* 29, 2860-2874

Amino-acid build-up around the DNA bases is spatially specific.

Arg distributions around adenine and guanine



Kono & Sarai (2001) "Structure-based prediction of DNA target sites by regulatory proteins." *Proteins* 35, 114-131

Atlas of Protein Side-chain Interactions

http://www.biochem.ucl.ac.uk/bsm/sidechains/index.html

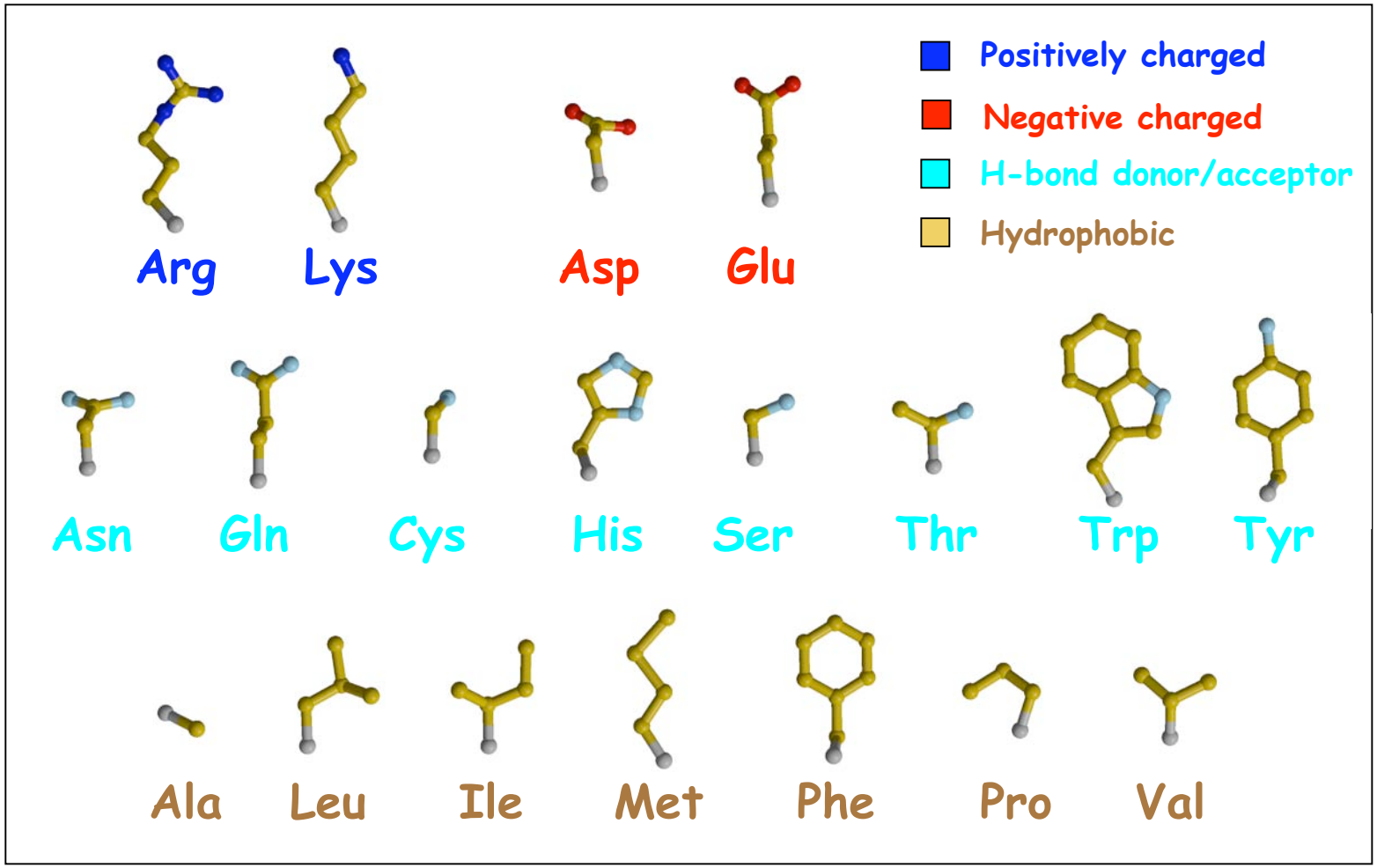
Sequence-dependent binding preferences of close interatomic contacts in protein-DNA complexes

DNA fragment	Sample size	Amino acid atom type			
		Cationic	Anionic	Nonpolar	Polar
Base	8240	1958	231	274	1417
A	**2234**	**275**	32	90	**510**
T	**2234**	**408**	20	**141**	**376**
G	**1886**	**1074**	4	26	**353**
C	**1886**	201	**175**	17	178
Phosphate	7352	2411	175	289	2237
Sugar	8240	986	87	267	1023
Total	23832	5355	493	830	4677

Number of close (≤ 3.4 Å) contacts between protein and DNA atoms in 239 protein-DNA complexes

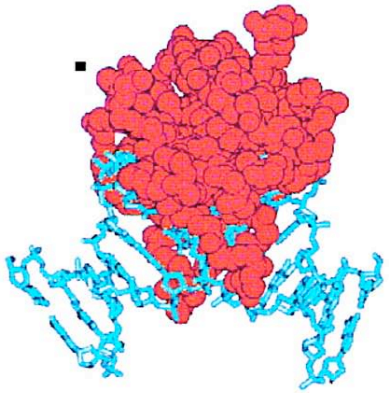
Yun Li

Amino-acid side groups and atoms color-coded by chemical make-up

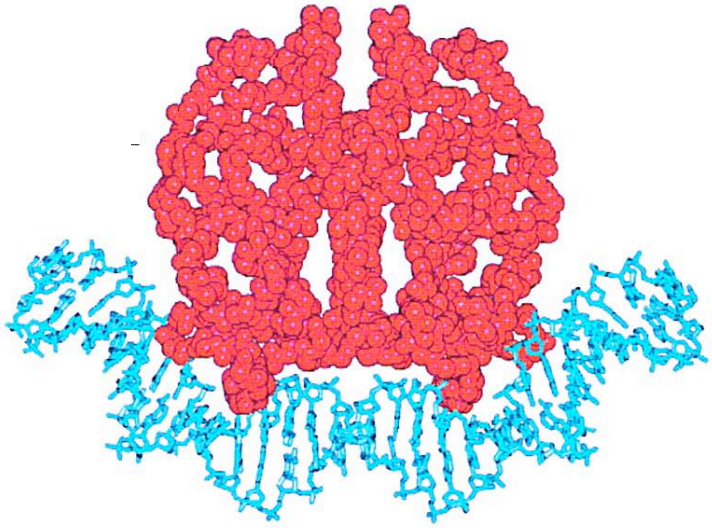


Charge neutralization and protein-induced DNA bending

Classes of DNA bending proteins



Class 1



Class 2

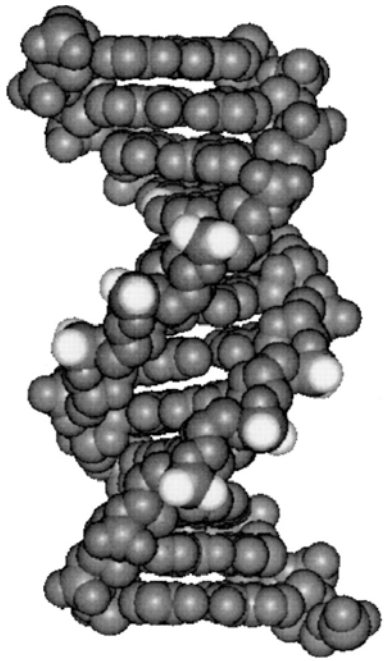
Class-1 bending proteins such as TBP bind the DNA minor groove, unwind DNA, and induce bending away from the protein-DNA interface by intercalation of hydrophobic-amino-acid side chains between base pairs.
Class-2 bending proteins such as CAP form complexes in which DNA bends toward the protein-DNA interface.

Williams & Maher (2000) "Electrostatic mechanisms of DNA deformation."
Ann. Rev. Biophys. Biomol. Struct. 29, 497-521

Phantom proteins designed to bend DNA

Sugar-base *i*–1
|
O3´
(–0.220)
|
O1P–P–O2P
(–0.830) (0.816) (–0.830)
|
O5´
(–0.220)
|
Sugar-base *i*
(0.284)

(Net charge –1)
Poltev *et al.* 1986



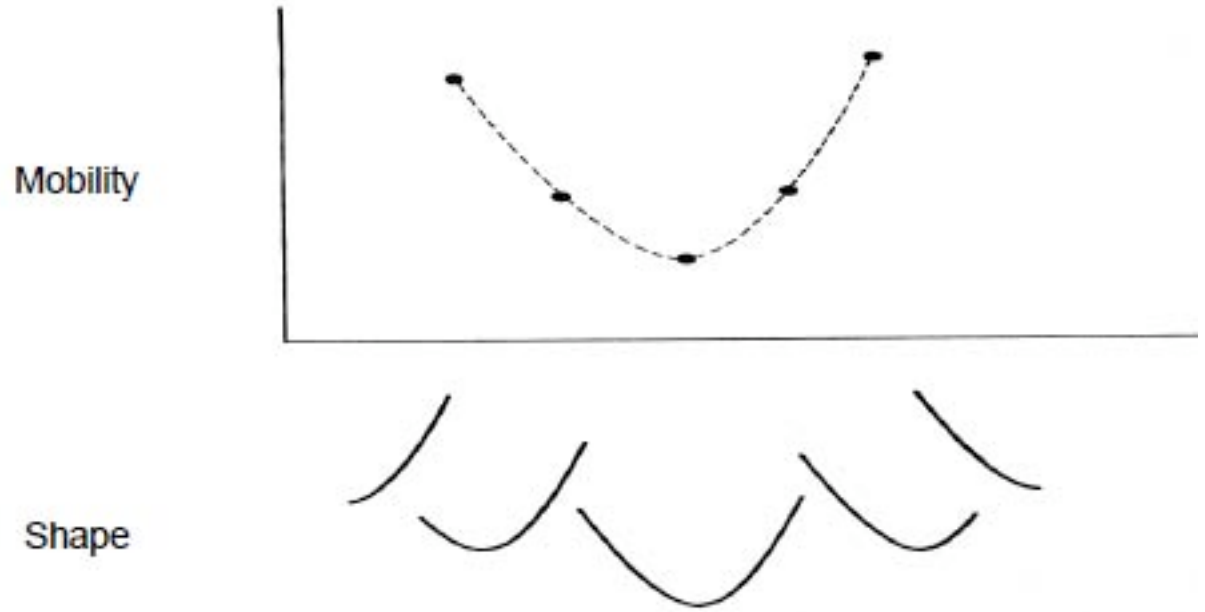
Sugar-base *i*–1
|
O3´
(–0.220)
|
Me–P–O2P
(0.013) (0.816) (–0.830)
|
O5´
(–0.220)
|
Sugar-base *i*
(0.284)

(Net charge –0.157)
Kosikov *et al.* 2001

Synthetic DNA duplexes in which selected phosphates are chemically neutralized by substitution of methylphosphonate analogs. *White spheres* indicate positions of methyl groups that neutralize consecutive phosphates across the minor groove.

Williams & Maher (2000) "Electrostatic mechanisms of DNA deformation."
Ann. Rev. Biophys. Biomol. Struct. 29, 497-521

Curved DNA molecules move more slowly than expected from their chain length on electrophoretic gels.

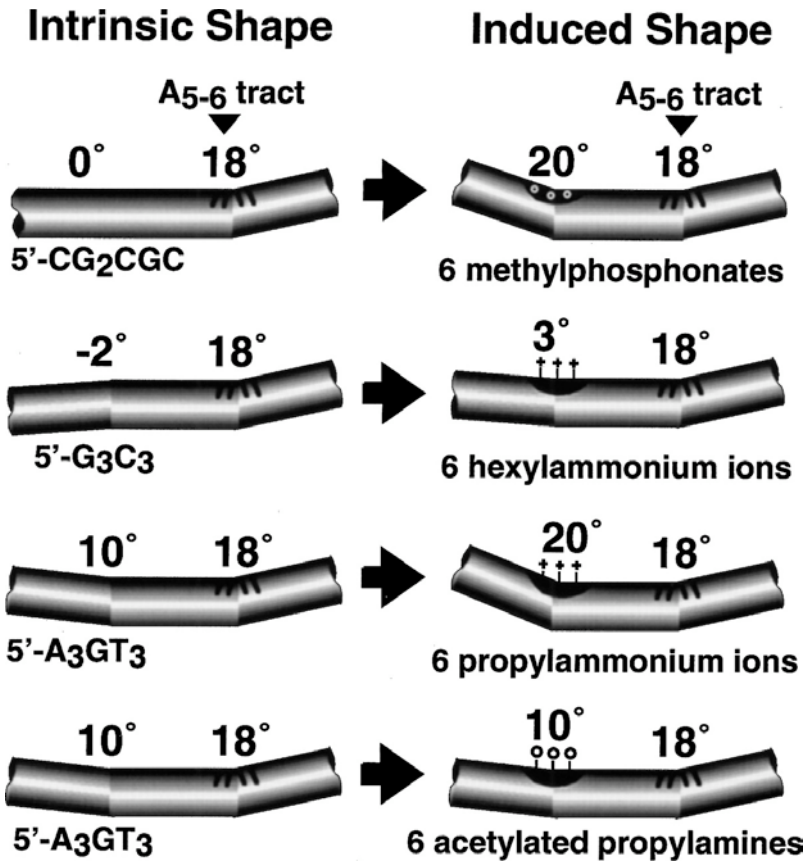


DNA molecules with naturally curved stretches located at different points along the chain sequence, i.e., ends vs. center, show different levels of mobility on polyacrylamide gels.

GGAATTCGAG CTCGCCCGGG GATCCGGCCT AAAATTCCAA CCGAAAATCG
CGAGGTTACT TTTTTGGAGC CCGAAAACCA CCCAAAATCA AGGAAAAATG
GCCAAAAAAT GCCAAAAAAT AGCGAAAATA CCCCGAAAAT TGGCAAAAAT
TAACAAAAAA TAGCGAATTT CCCTGAATTT TAGGCGAAAA AACCCCCGAA

Phased tracts of A's (such as the above from the kinetoplast DNA of *Crithidia fasiculata*) make DNA so strongly curved that chain fragments less than 200 bp easily close into a circle .

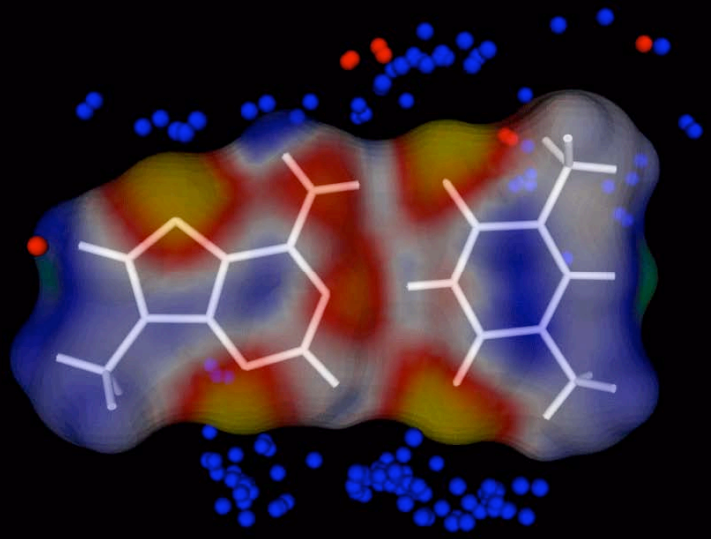
DNA bending induced by phantom proteins



Proteins simulated by methylphosphonate substitution or appending cations on propyl or hexyl tethers. Intrinsic DNA shapes of the indicated sequences shown as cylinders at left, including the position of reference A_{5-6}-tracts. Induced shapes shown at right. Degree of bending deduced from gel mobilities.

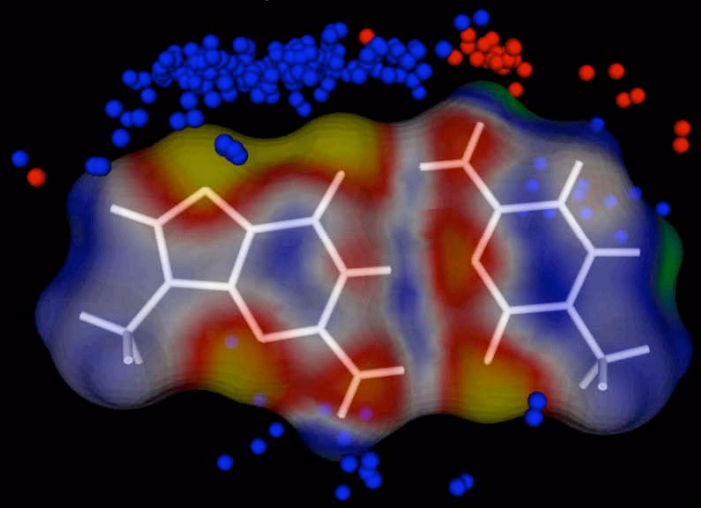
Williams & Maher (2000) "Electrostatic mechanisms of DNA deformation."
Ann. Rev. Biophys. Biomol. Struct. 29, 497-521

The sequence-dependent build-up of charged amino acids around the DNA bases is a type of localized nucleotide charge neutralization.



A·T minor groove edge

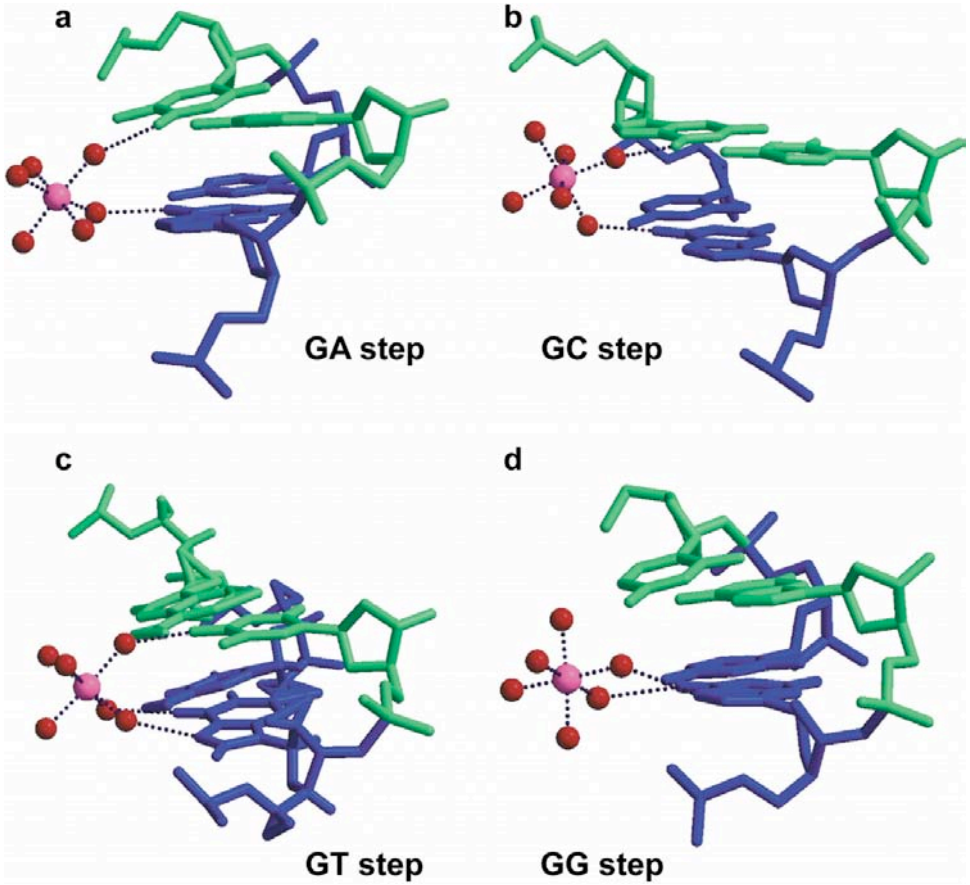
G·C major groove edge



Arg/Lys_N+
Asp/Glu_O–

Sites of preferred contact with DNA reflect the intrinsic electronic structure of the base pairs.

Solvated cations similarly neutralize the major-groove edge of guanine.

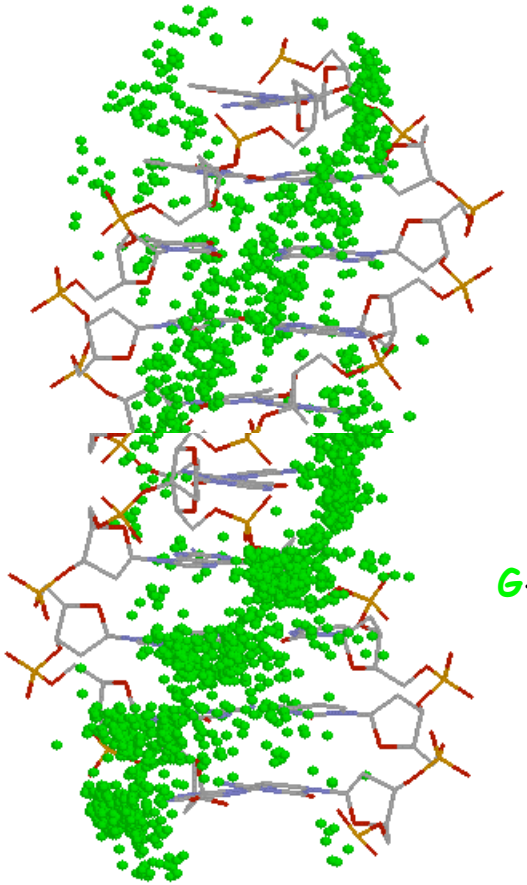


Mg^{2+} bound to guanine-containing steps in B-form DNA. G shown in all cases on the front side of the lower base pair. Ions taken from the following structures: (*a*) BD0037, (*b*) BD0007, (*c*) BD0033, and (*d*) BD0033.

Subirana & Soler-López (2003) "Cations as hydrogen bond donors." Ann. Rev. Biophys. Biomol. Struct. 32, 27-45

The build-up of cationic species on different edges of A·T vs. G·C base pairs leads to the accumulation charge on one face of the double-helix.

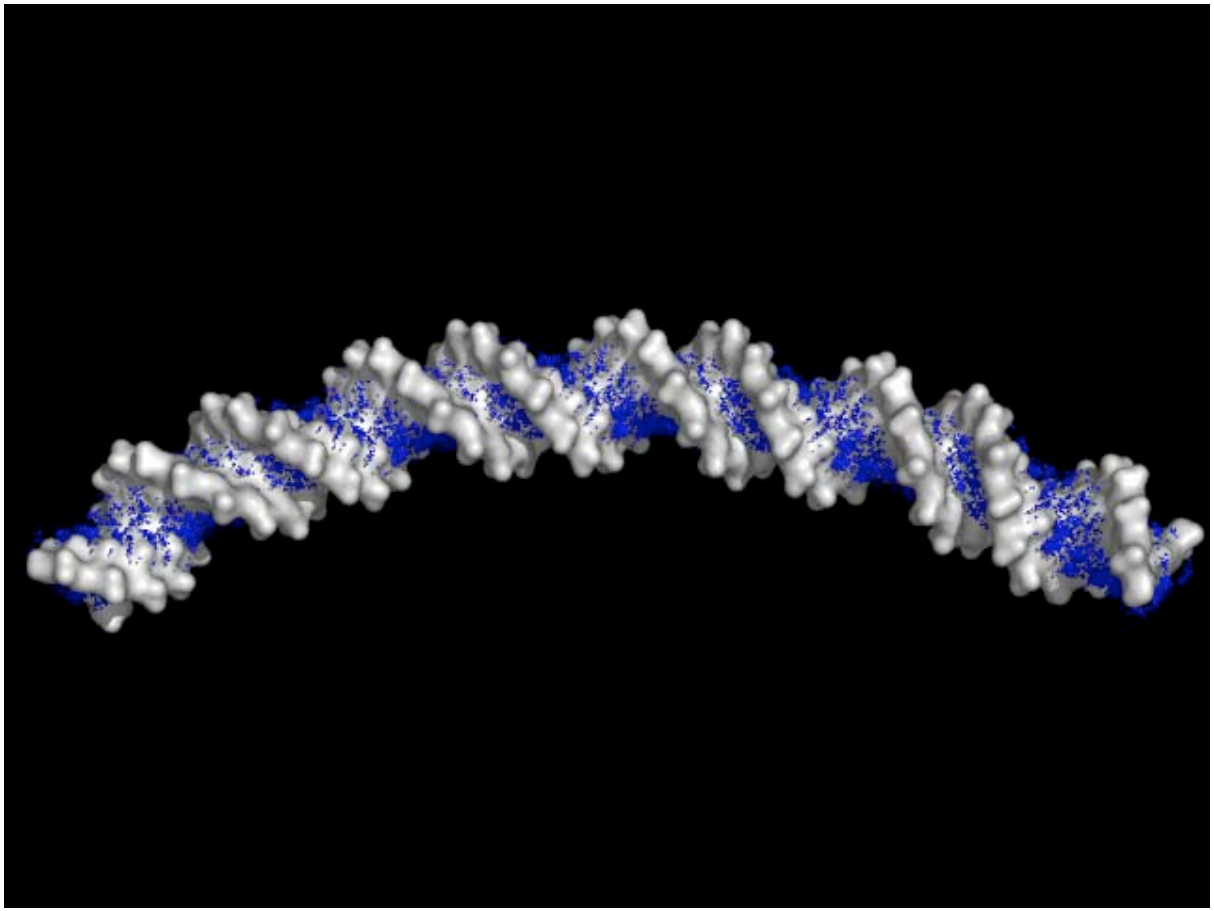
A·T minor-groove



G·C major-groove

poly dA_5G_5·poly dC_5T_5

DNA sequences with excess charge on one side of the double helix will bend toward the neutralized face.



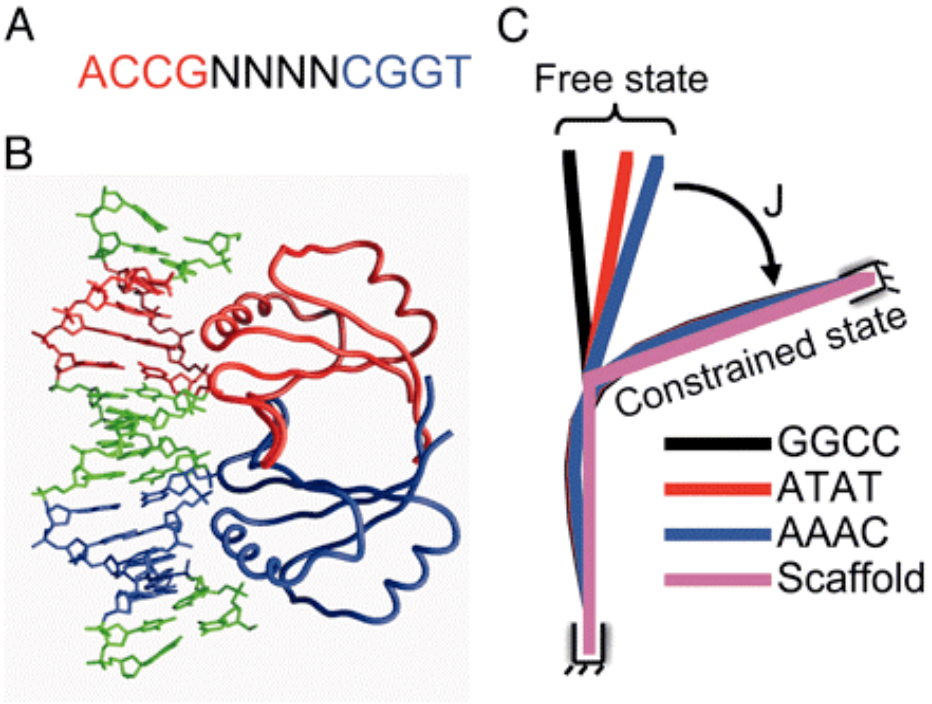
The neutralization of phosphates by a localized excess of cationic charge curves the DNA.

The curvature, in turn, facilitates the wrapping of DNA on proteins and the long-range recognition of specific sequences.

Indirect recognition

DNA sequence discrimination may be indirect, involving the sequence-dependent structure or deformability of double helix.

- The E2 protein from human papillomavirus discriminates a spacer insert in its target sequence without touching the bases.
- Indirect recognition: The recognition of sequence-dependent structural features of DNA, such as intrinsic curvature, deformability, or hydration patterns.

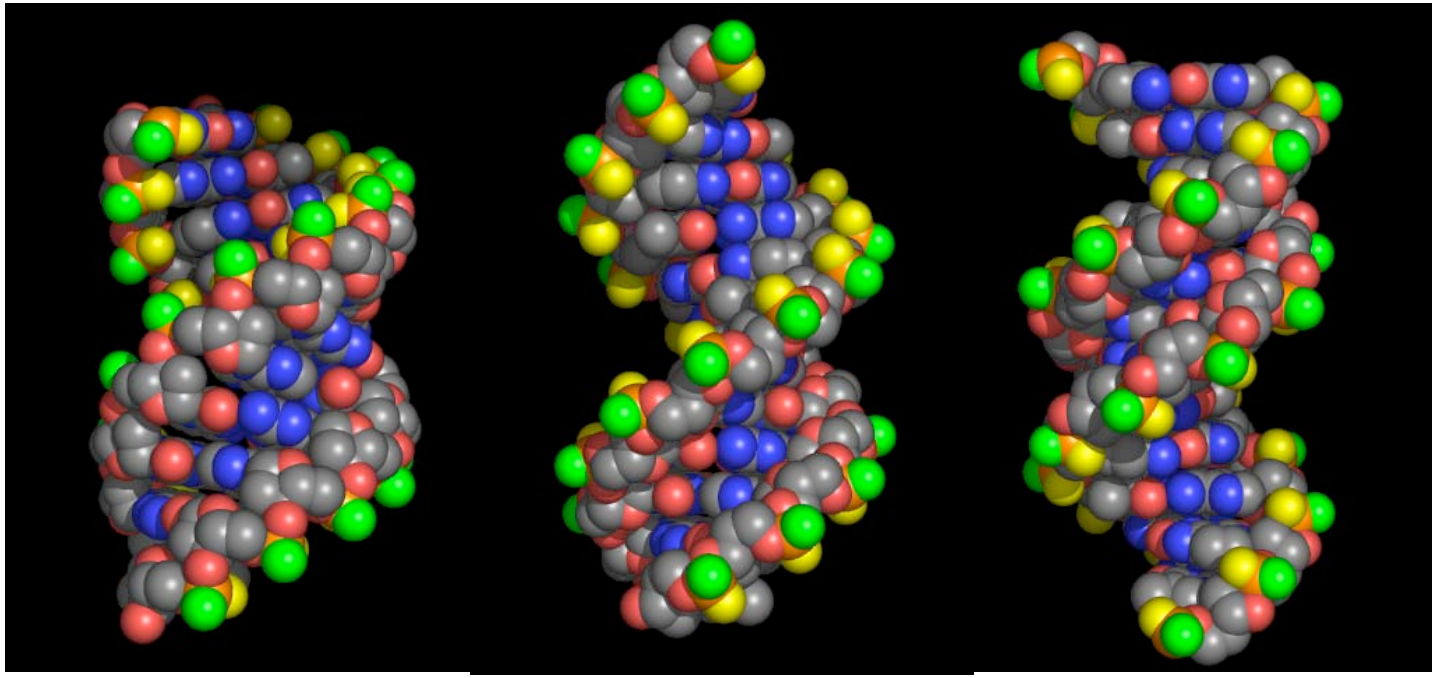


Zhang *et al.* (2004) "Predicting indirect readout effects in protein-DNA interactions." *Proc. Natl. Acad. Sci. USA* 101, 8337-8341

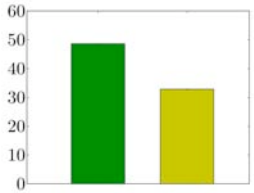
Shape complementarity

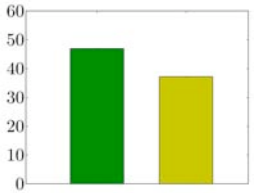
The exposure of the DNA phosphates depends upon helical state,

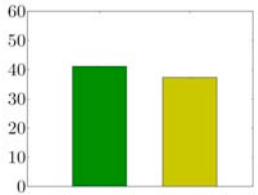
A DNA B DNA C DNA



Solvent accessible area (Å)



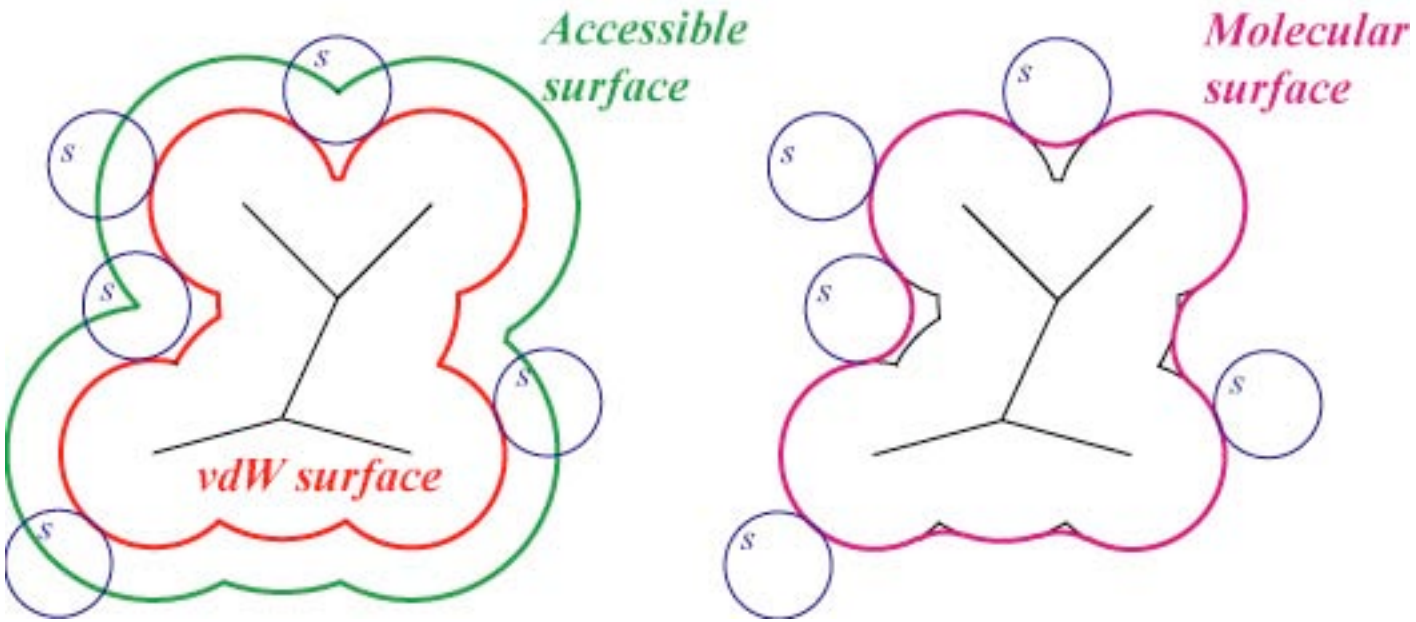




O1P

O2P

Schematic illustration of molecular surfaces



- vdW surface (red) - union of balls representing all atoms, with radii set to the vdW radii.
- accessible surface (green) - surface generated by the center of a sphere rolling on the vdW surface. The radius of this sphere is usually set to 1.4 Angstroms, the radius of a water molecule.
- molecular surface (magenta) - lower envelope generated by the rolling sphere. It differs from the vdW surface in that some areas are inaccessible to the rolling sphere

Image from the UC Davis Genome website: http://koehllab.genomecenter.ucdavis.edu

Surface accessiblity can also be detected by "footprints" of chemical reactivity.

Color-coded representation of the solvent-accessible surface of IHF-bound DNA revealed by the hydroxyl radical footprint in solution and superimposed on the crystal complex (PDB_ID: 1ihf): **green**, **light blue**, and **dark blue** indicate mild, moderate, and strong protection from cleavage; **yellow**, **orange**, and **red** indicate mild, moderate, and strong enhancement of cleavage.

Khrapunov *et al.* (2006) "Binding then bending: A mechanism for wrapping DNA."
Proc. Natl. Acad. Sci., USA 51, 19217-19218

The surfaces of interacting molecules are complementary.

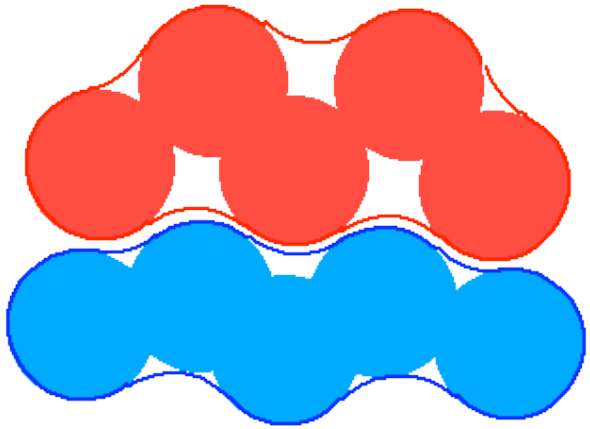
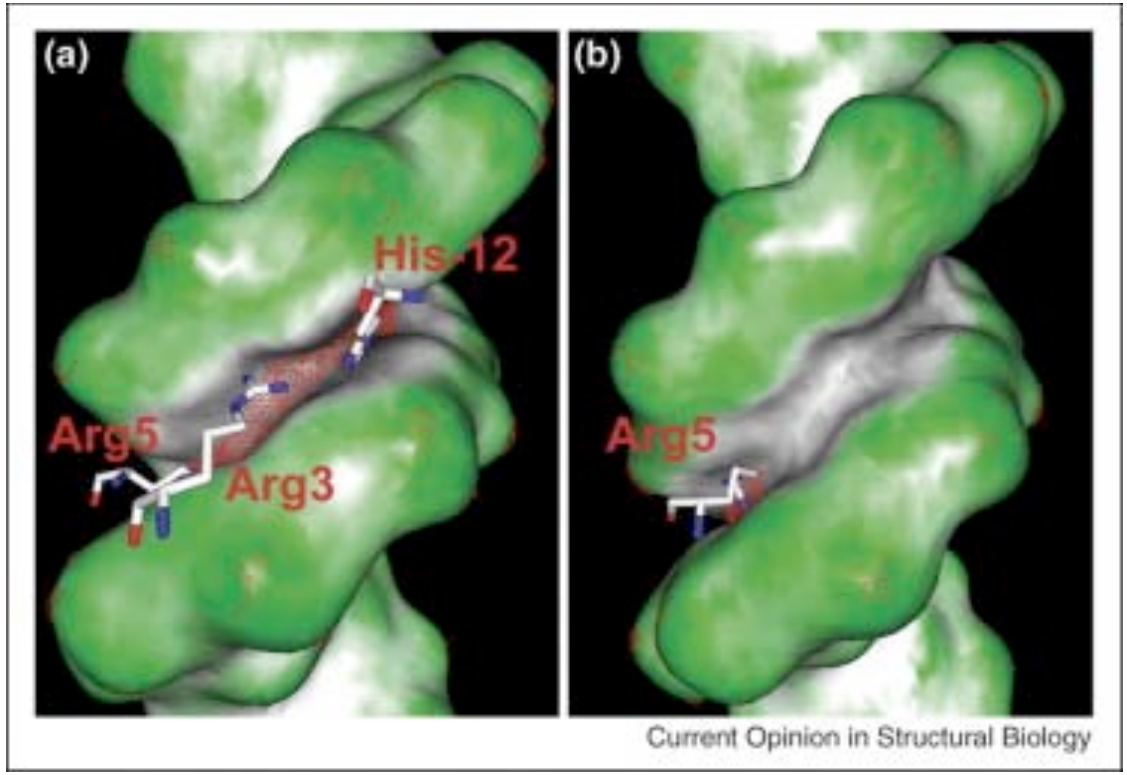


Image from the Network Science website: http://www.netsci.org

Recognition of minor-grove shape and electrostatic potential by a Hox homeodomain



Scr bound to (a) its specific recognition sequence and (b) the Hox consensus sequence (PDB_IDs 2r5z, 2r5y). Note the extended narrow minor groove, which binds His-12, Arg3, and Arg5 in (a) vs. the short narrowed region that binds Arg5 in (b). Also note the more negative (red) electrostatic potential in (a) vs. (b).

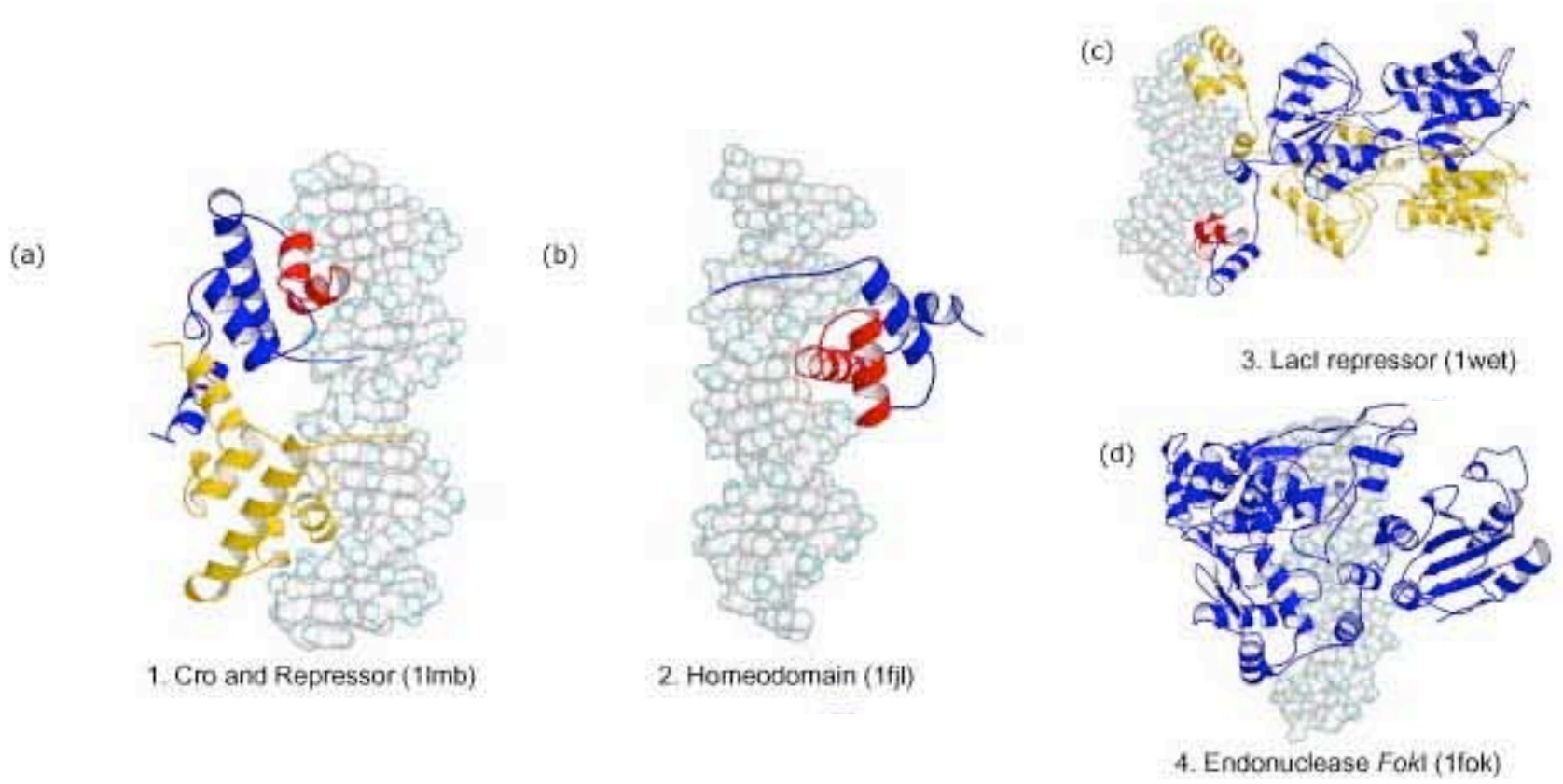
Rohs *et al.* (2009) "Nuance in the double-helix and its role in protein-DNA recognition." *Curr. Opin. Struct. Biol.* 19, 171-177

Summary points from the DNA perspective

- Proteins typically recognize DNA sequence via direct hydrogen bonding or van der Waals interactions with the constituent nucleotides.
- Concomitant conformational changes in DNA — sequence-dependent kinking, helical dislocation, untwisting, intercalation, etc. — contribute to the fit of protein against DNA.
- The deformability in DNA is essential at both the global and local levels, serving as a potential long-range signal for molecular recognition as well as accommodating the local distortions of the double helix induced by tight binding.
- The conformational recognition of DNA sequence is often referred to as indirect readout.

Protein perspective

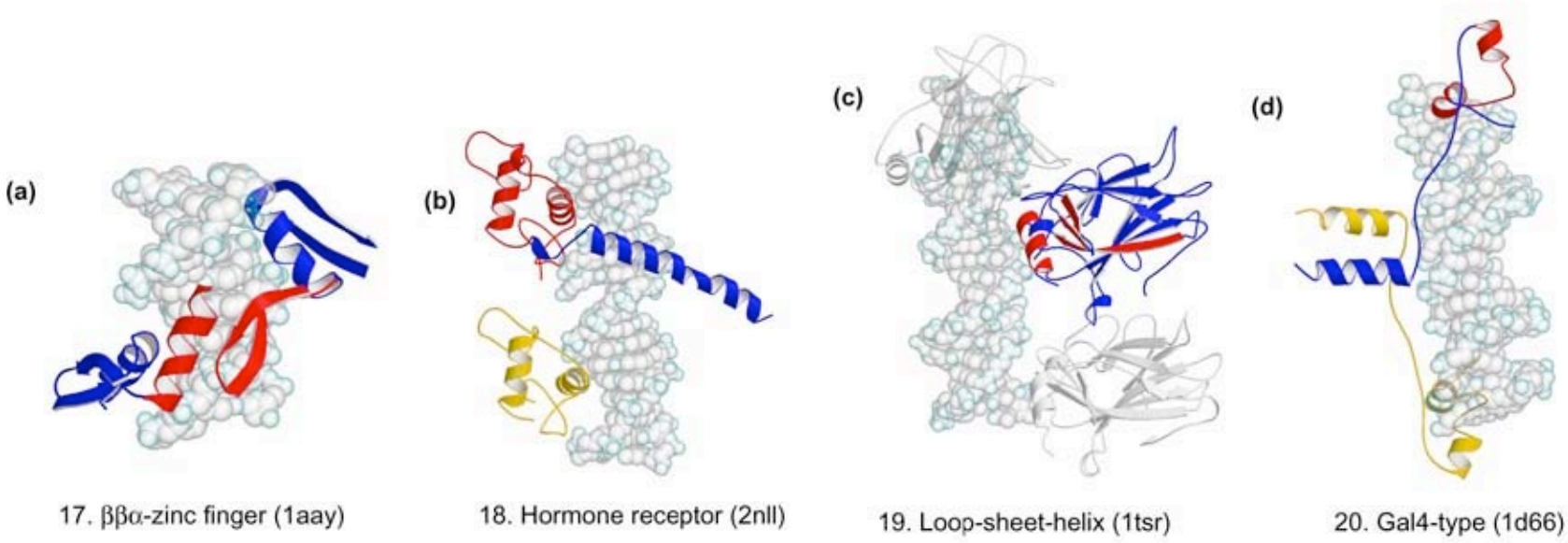
Proteins recognize DNA via various independently folded binding domains.



The helix-turn-helix (HTH) motif — roughly 20 amino acids folded into two roughly perpendicular α-helices linked by a β-turn or loop — is used by transcription regulators and enzymes of both prokaryotes and eukaryotes typically to bind the major-groove edges of the DNA base pairs. The linker and non-recognition α-helix tend to interact with the sugar-phosphate backbone. Winged HTH proteins (such as the homeodomain) contain a third α-helix.

Luscombe *et al.* (2000) "An overview of the structures of protein-DNA complexes." *Genome Biol.* 1, 1-37

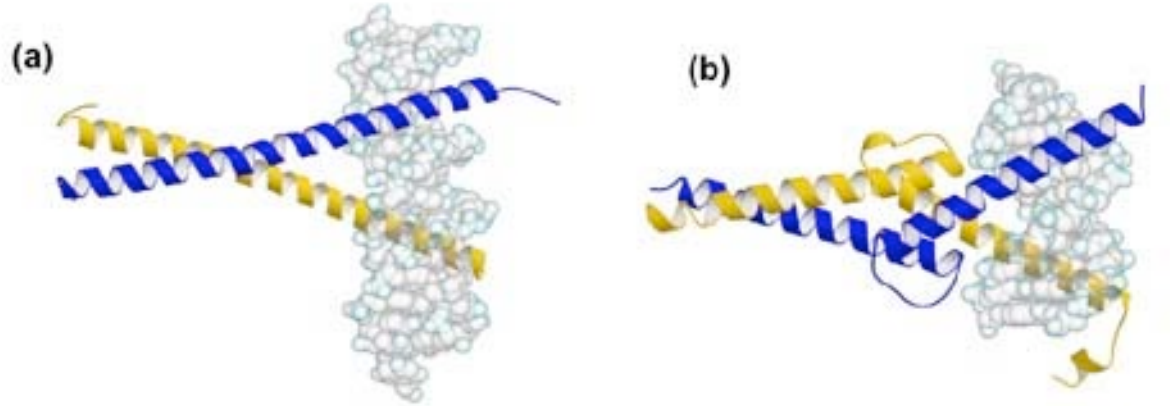
Proteins recognize DNA via various independently folded binding domains.



Zinc-coordinating proteins — the largest group of transcription factors in eukaryote genomes — entail the tetrahedral coordination of 1-2 zinc ions with conserved cystein and histidine residues. The motif is used in protein-protein interactions as well as DNA binding. Zinc fingers, made up of a recognition α-helix and a 2-strand β-sheet and constituting the largest family in thie group, recognize a triplet of DNA base pairs.

Luscombe *et al.* (2000) "An overview of the structures of protein-DNA complexes." *Genome Biol.* 1, 1-37

Proteins recognize DNA via various independently folded binding domains.



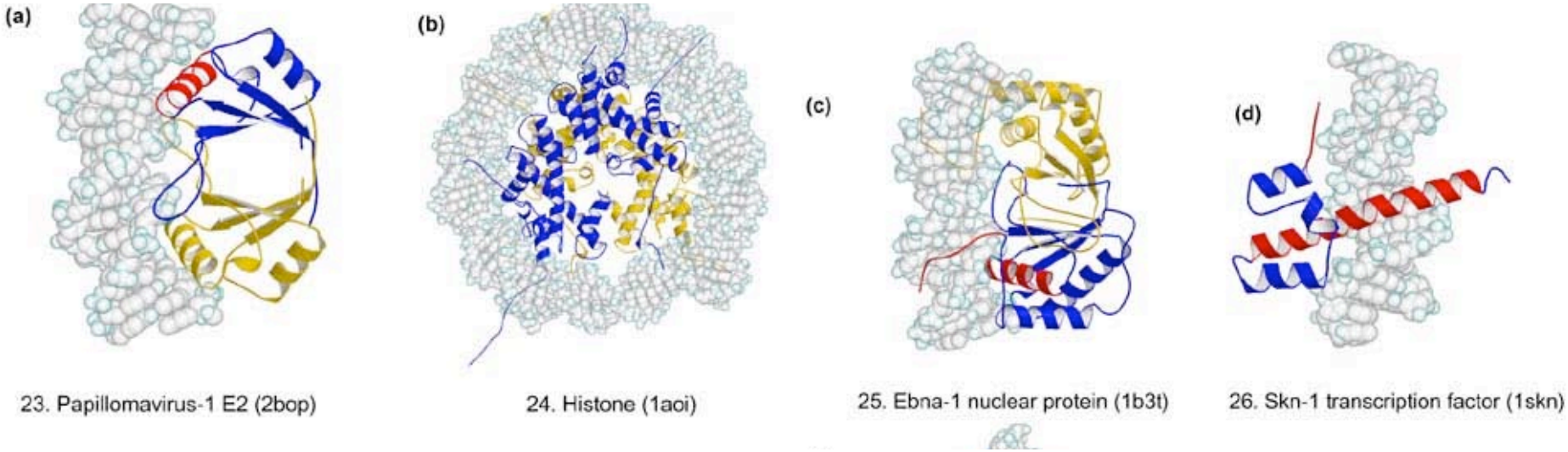
21. Leucine zipper (2dgc)

22. Helix-loop-helix (1am9)

Zipper-type proteins — The **leucine zipper** (bZIP) contains an a-helix with a leucine at every 7th amino acid. The leucines act as the teeth of a zipper that allows dimerization of two proteins. Basic amino acids bind to the sugar-phosphate backbone while the helices fit in the major grooves on opposite sides of the duplex. The **helix-loop-helix** motif includes two α helices connected by a loop. One helix, which istypically smaller, dimerizes by folding and packing against another helix. The larger helix typically contains the DNA binding regions

Luscombe *et al.* (2000) "An overview of the structures of protein-DNA complexes." *Genome Biol.* 1, 1-37

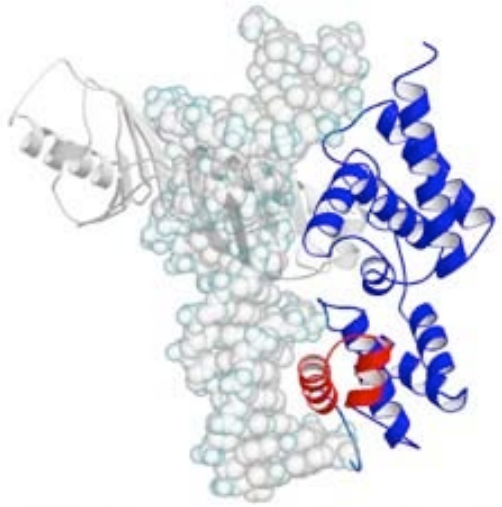
Proteins recognize DNA via various independently folded binding domains.



Other α-helix proteins — These proteins fall into seven distinct families with very different folding patterns and a variety of functions. All use α-helices as the main binding motif.

Luscombe *et al.* (2000) "An overview of the structures of protein-DNA complexes." *Genome Biol.* 1, 1-37

Proteins recognize DNA via various independently folded binding domains.

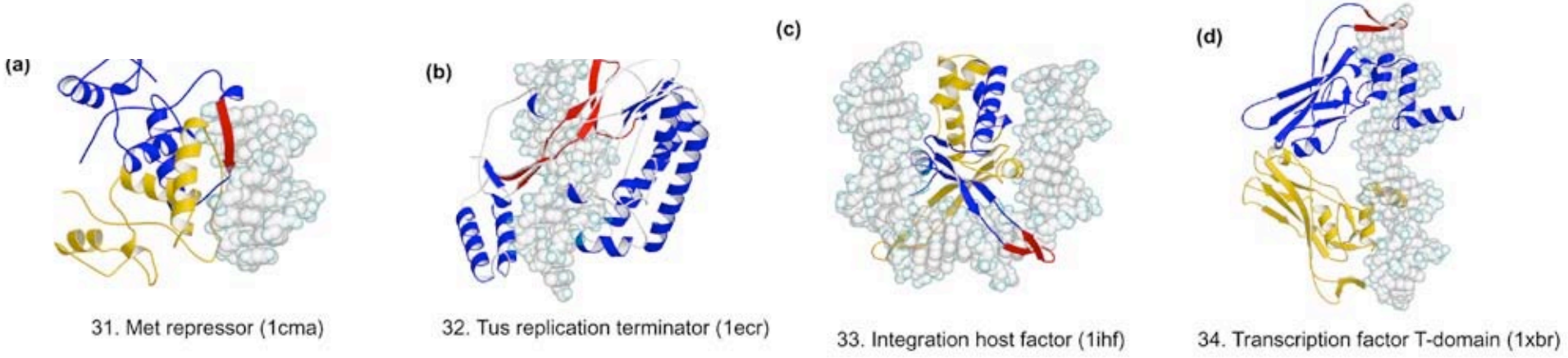


30. TATA box-binding family (1ytb)

β-sheet proteins —TBP, the single member of this group, uses β-strands as recognition and binding motifs.

Luscombe *et al.* (2000) "An overview of the structures of protein-DNA complexes." *Genome Biol.* 1, 1-37

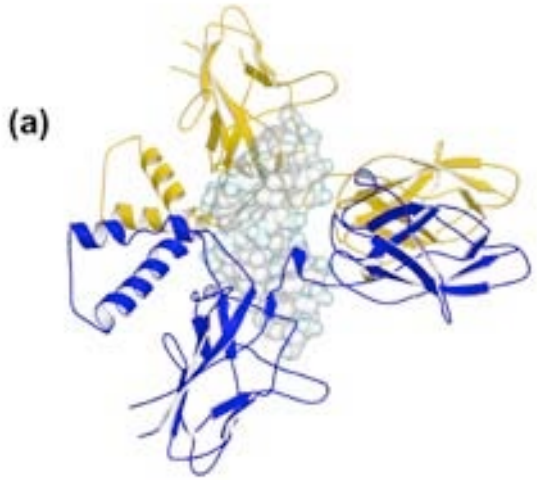
Proteins recognize DNA via various independently folded binding domains.



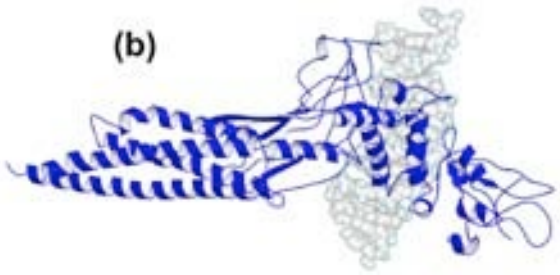
β-hairpin/ribbon proteins —This group of six proteins use small 2- and 3-stranded β-sheets or hairpin motifs to bind the DNA major or minor grooves.

Luscombe *et al.* (2000) "An overview of the structures of protein-DNA complexes." *Genome Biol.* 1, 1-37

Proteins recognize DNA via various independently folded binding domains.



37. Rel homology region (1a3q)

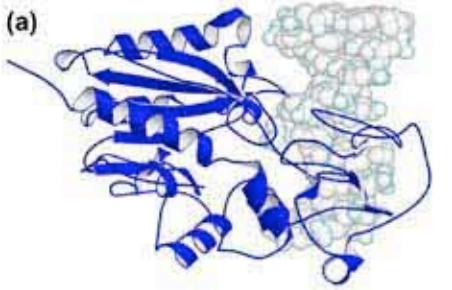


38. Stat family (1bf5)

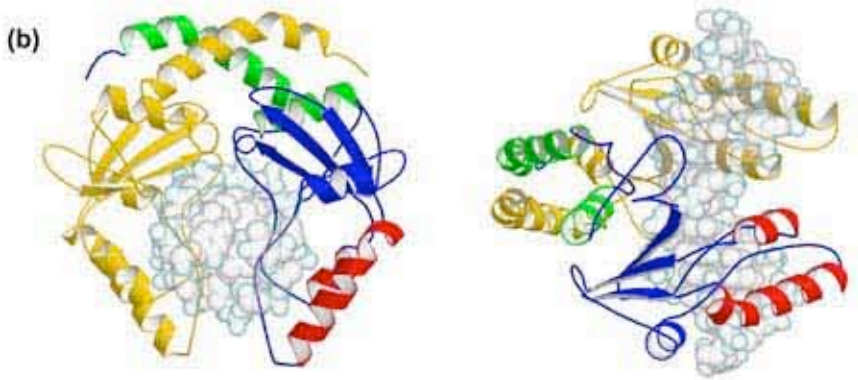
Other proteins — Some types of non-enzymatic proteins employ no well-defined secondary structural motif for DNA recognition. The above examples function as dimers, use multi-domain subunits, and envelop their DNA binding partner.

Luscombe *et al.* (2000) "An overview of the structures of protein-DNA complexes." *Genome Biol.* 1, 1-37

Proteins recognize DNA via various independently folded binding domains.



39. Methyltransferase family (6mht)



40. Endonuclease PvuII family (3pvi)

Enzymes — This group of protein is based on biological function rather than structure. The DNA-binding regions do not fall into simple structural categories. The proteins use various combinations of α-helices, β-strands, and loops to recognize DNA. Many enzymes contain three domains: a DNA-recognition domain that 'reads' sequence; a catalytic domain with the enzyme active site; where applicable, a dimerization domain. The bound DNA is often highly deformed.

Luscombe *et al.* (2000) "An overview of the structures of protein-DNA complexes." *Genome Biol.* 1, 1-37

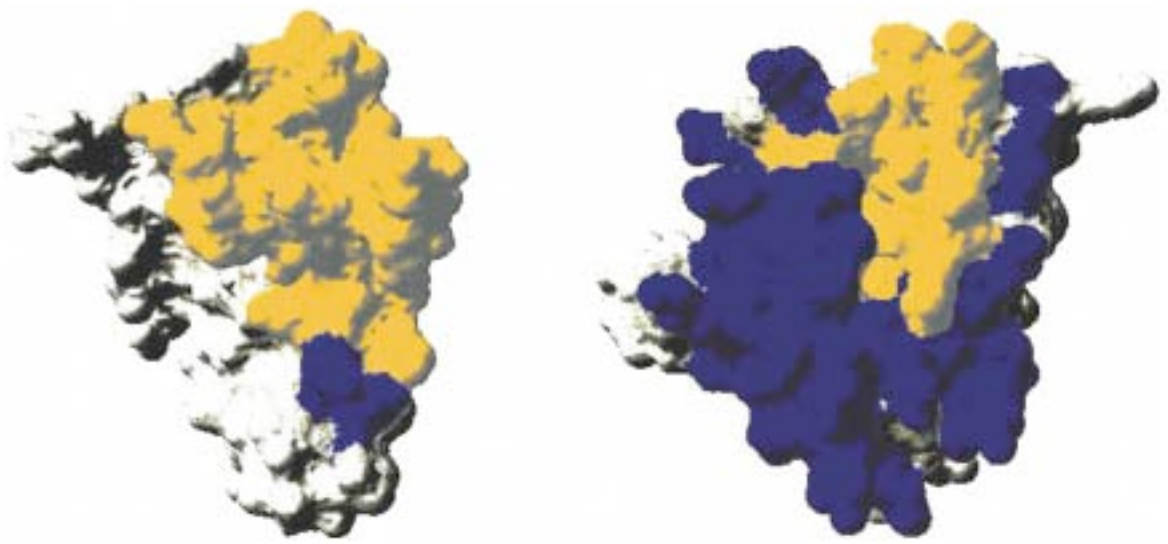
Protein-centric overview of PDB structures
http://www.ebi.ac.uk/pdbsum/

Summary of DNA-binding protein structural families

http://www.biochem.ucl.ac.uk/bsm/prot_dna/prot_dna_cover.html

Is it possible to predict whether a protein will bind DNA or RNA?

The DNA-binding property of some proteins can be predicted with high accuracy from the structural and sequential properties of the large, positively charged electrostatic patches characteristic of known protein-DNA complexes.

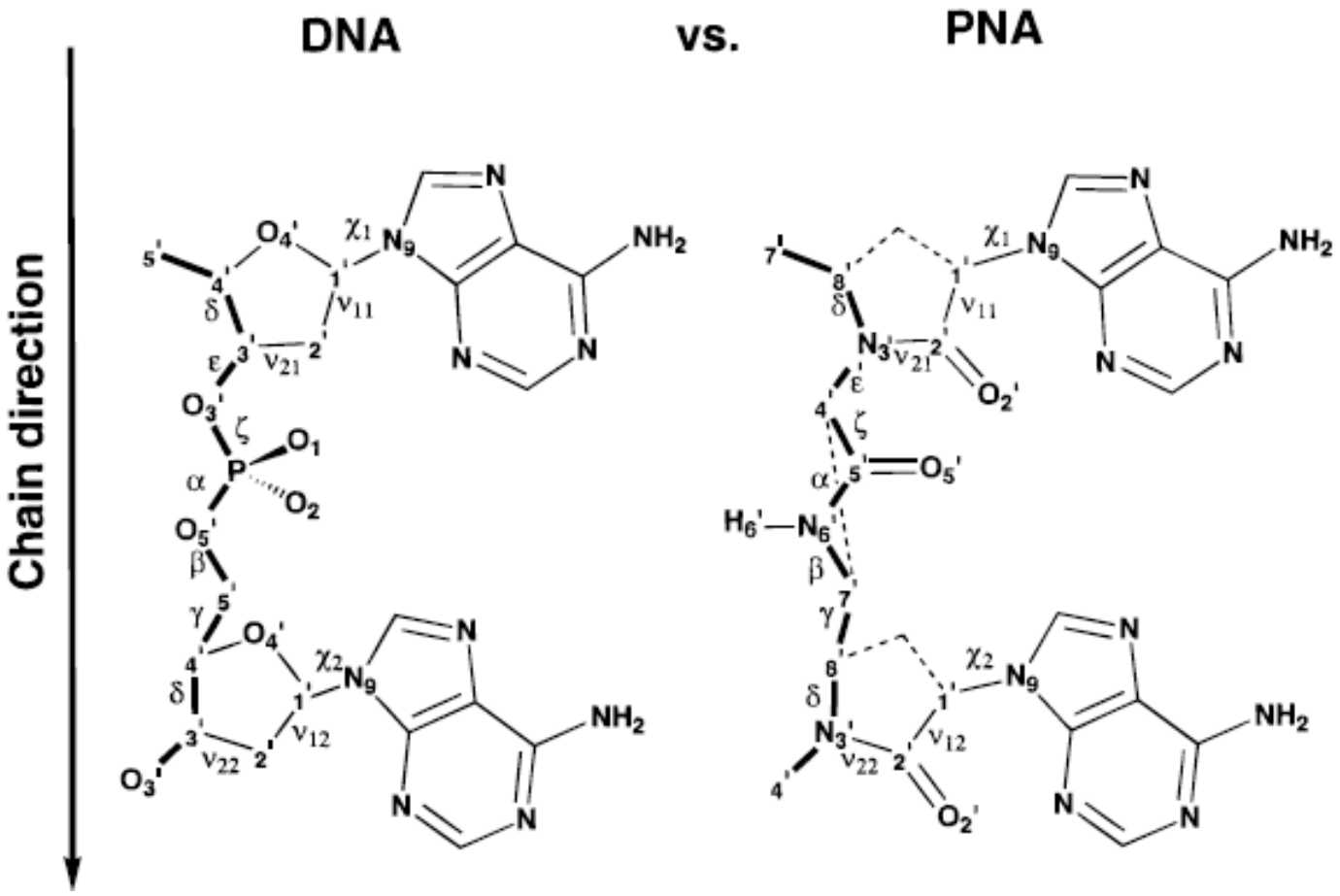


Surface electrostatic patches of two different proteins: (left) MCM1, a DNA-binding protein (PDB_ID: 1mnm); (right) cytochrome c3, a non-nucleic-acid binding protein (PDB_ID: 1cot). The former protein retains a large percentage of its positive patch even after the charge on lysine is removed (yellow area). The patch of cytochrome c2 shrinks dramatically.

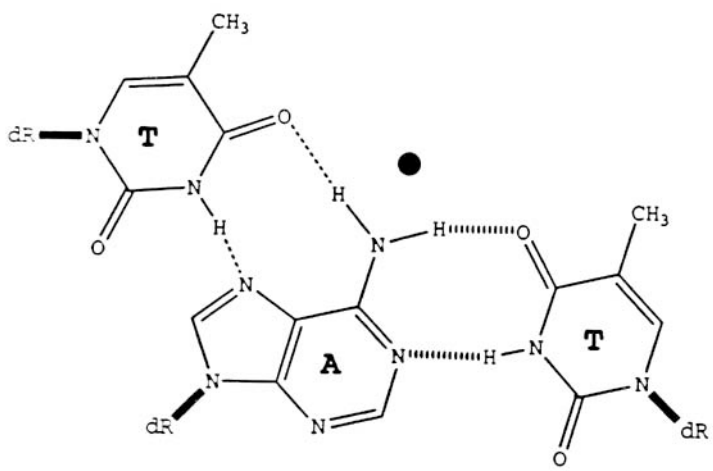
Stawiski *et al.* (2003) "Annotating nucleic-acid-binding function based on structure." *J. Mol. Biol.* 326, 1065-1079

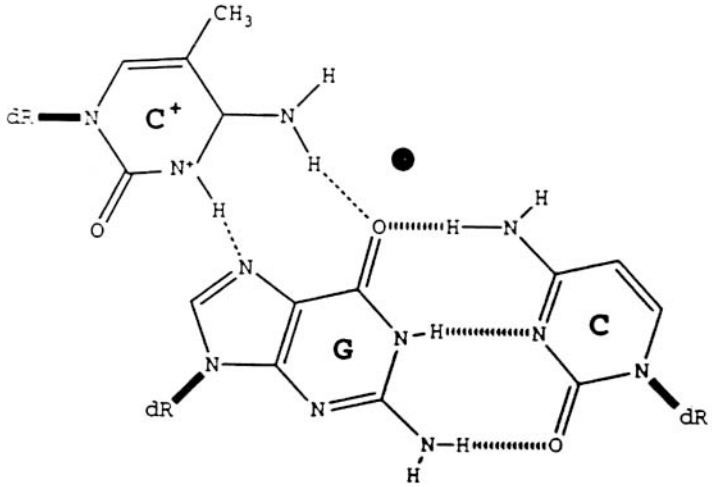
PNA-DNA recognition

Peptide nucleic acids (PNA) combine the information-storage properties of DNA with the chemical stability of a protein-like backbone.



PNA recognizes DNA sequence via major-groove Hoogsteen base pairing.





T:A·T and 5MeC$^+$: G·C triplets. Hoogsteen pairing denoted by dashed lines.

Schematic of PNA bound in the DNA major groove

Nielsen (2008) "Triple helix: designing a new molecule of life." *Sci. Amer.* 299, 64-71