Biochemistry

© Copyright 1991 by the American Chemical Society Volume 30, Number 48 December 3, 1991

Perspectives in Biochemistry

DNA Binding Specificity of Homeodomains[†]

Allen Laughon

Laboratory of Genetics, University of Wisconsin, Madison, Wisconsin 53706

Received September 9, 1991; Revised Manuscript Received October 10, 1991

In the eight years since the discovery of the homeodomain (McGinnis et al., 1984a,b; Scott & Wiener, 1984), this highly conserved DNA-binding domain has been found in dozens of transcription factors (Scott et al., 1989; Rosenfeld, 1991). Genetic and molecular studies of *Drosophila* have identified more than 20 homeodomain proteins that regulate cell determination within specific developmental pathways (Akam, 1987; Hayashi & Scott, 1990). Vertebrate homologues have been identified for almost every known *Drosophila* homeodomain, providing a starting point for directed mutational analysis of vertebrate development (Kessel & Gruss, 1990).

The regulatory function of a homeodomain protein derives from the specificity of its interactions with DNA and presumably with components of the basic transcriptional machinery such as RNA polymerase or accessory transcription factors. Although little is known about the second type of specificity, the last several years have witnessed important advances in our understanding of how homeodomains contact DNA. The most important contributions have been the determination of three-dimensional structures for two homeodomain–DNA complexes (Otting et al., 1990; Kissinger et al., 1990). These studies confirmed previous predictions that homeodomains utilize a helix–turn–helix (HTH) fold to contact DNA in the major groove (Laughon & Scott, 1984; Shepherd et al., 1984) but together with genetic studies (Hanes & Brent, 1989; Treisman et al., 1989) revealed that homeodomain HTH–DNA contacts are different from the prokaryotic HTH paradigm. In addition, homeodomains utilize an N-terminal arm to contact DNA in the minor groove. This minor-groove contact is sequence-specific and contributes substantially to the high affinity of homeodomains for DNA (Affolter et al., 1990; Percival-Smith et al., 1990; Florence et al., 1991). Here, I review these recent findings and consider their implications for the regulatory specificity of homeodomain proteins.

Homeodomains Are Helix–Turn–Helix DNA-Binding Domains

Prior to the discovery of homeodomains, crystal structures had been solved for several prokaryotic repressor proteins (Pabo & Sauer, 1984). Although the overall structures of these proteins are quite different, each contains a nearly identical helix–turn–helix (HTH) DNA-binding motif. The second helix, or "recognition helix", makes base-specific contact with DNA in the major groove, while helix 1 lies across helix 2. HTH structures contain characteristic hydrophobic amino acids at the interface between the two helices and at the turn between them. The C-terminal halves of homeodomains contain a similar motif. The amino acids corresponding to the recognition helix of the HTH motif are highly conserved among a large number of homeodomains, the first indication that many homeodomain proteins have closely related DNA-binding specificities. The discovery of homeodomains was soon followed by experiments demonstrating sequence-specific binding of homeodomains to DNA (Desplan et al., 1985).

The structures of homeodomain–DNA complexes have been determined using NMR spectroscopy for the Antennapedia (*Antp*) homeodomain (Otting et al., 1988, 1990; Qian et al. 1989) and X-ray crystallography for the engrailed (*en*) homeodomain (Kissinger et al., 1990). The two homeodomain–DNA structures are very similar, utilizing a HTH fold to make base-specific contact with 3–4 bp in the major groove and an N-terminal arm to contact 2 bp in the minor groove (Figure 1). The homeodomain contains three helices, with helices 2 and 3 forming the HTH. Helix 1 lies across helix 3 parallel to helix 2. The entire structure is held together by a pocket of hydrophobic amino acids at the interface between the three helices. In both structures, the homeodomain is bound to DNA as a monomer. The equilibrium binding constants for high affinity sites have been measured to be about 10^{-9} M for the *Antp* homeodomain (Affolter et al., 1990) and (2–60) × 10^{-11} M for the fushi tarazu (*ftz*) homeodomain, a close relative of *Antp* (Percival-Smith et al., 1990; Florence et al., 1991).

[†] Supported by grants from the American Cancer Society and the Milwaukee Foundation's Shaw Scholars Program. Laboratory of Genetics Paper No. 3212.

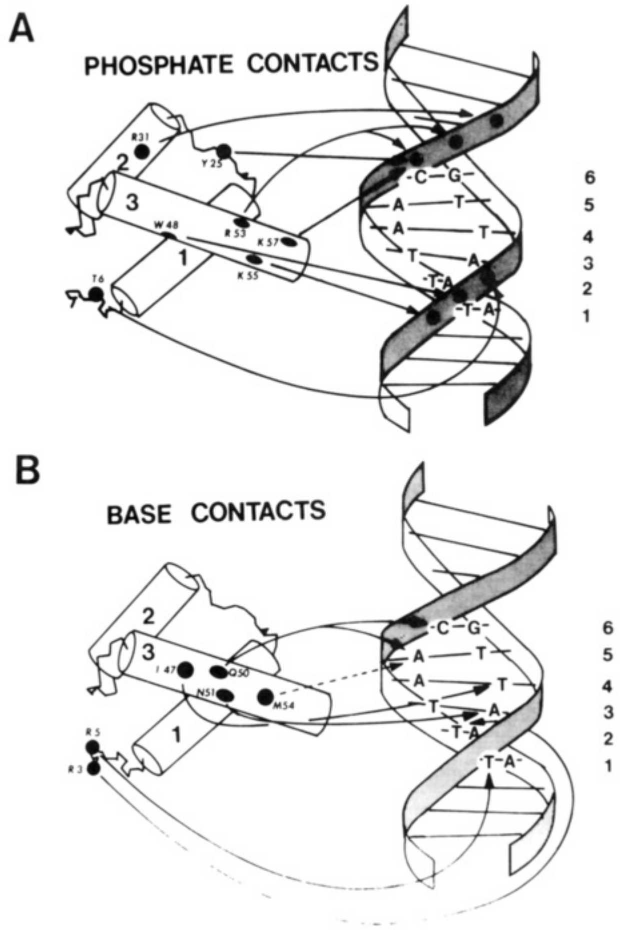

FIGURE 1: Homeodomain–DNA contacts. (A) Phosphate contacts. Lines with arrows connect amino acids with the phosphates (shown as solid circles) that they contact in the *en* complex. (B) Base contacts. Lines with arrows connect amino acids with the specific bases they contact in the *en* or *Antp* structures; see text for details. The drawings of homeodomain and DNA are adapted from Kissinger et al. (1990) for the *en* homeodomain–DNA complex, except that Met 54, corresponding to *Antp*, is shown instead of Ala 54 of *en*. The three α helices are numbered.

Under identical conditions, the affinity of the *Antp* homeodomain for specific sites is about 10 000-fold higher than the affinities of two prokaryotic repressor monomers for their corresponding high-affinity sites (Affolter et al., 1990). As shown in Table I, part of this large difference is due to the N-terminal arm, deletion of which reduces *ftz* binding affinity over 100-fold (Percival-Smith et al., 1990). The homeodomain also makes more extensive contact with backbone phosphates (Figure 1A) than do prokaryotic repressors (Jordan & Pabo, 1988; Aggarwal et al., 1988). A mutation altering one of the phosphate contact residues (Arg 53 → Ala) reduces activity of the *bcd* homeodomain 40–100-fold (Table I; Hanes & Brent, 1991). In addition, the recognition helix is longer in homeodomains than in prokaryotic recognition helices and contacts the major groove nearer its C-terminus as opposed to the N-terminal contact made by prokaryotic recognition helices. The C-terminal end of the 17 amino acid *en* helix 3 forms a separate helix 4 in *Antp* (Qian et al. 1989), and in both proteins, it contains three basic amino acids that contact phosphates on the DNA backbone.

Homeodomains have been found in invertebrates, vertebrates, fungi, and plants. A compilation of published homeodomain sequences is presented in Figure 2 and these can be divided up into 25 distinct homeodomain groups. Most groups contain both vertebrate and invertebrate members, evidence that a diverse array of homeodomains had evolved specialized structures prior to the radiation of metazoans.

Table I: Contributions of Homeodomain Base and Phosphate Contacts to DNA-Binding Affinity

amino acid	contact	mutation	*x*-fold reduction in binding affinity or transactivation activity
Arg 53	phosphate	*bcd* Arg 53 → Ala 53	40[a]
Thr 6	phosphate	*pit-1* Thr 6 → Ala 6	0[b]
		pit-1 Thr 7 → Ala 7	0[b]
Arg 3	T-1	T-1 → G (*ftz*)	8.9[c]
		T-1 → A	7.6[c]
Arg 5	T-2′	T-2′ → C	20[c]
		T-2′ → G	19[c]
		T-2′ → A	3.4[c]
Arg 3, Arg 5, Thr 6	T-1 + T-2′	*ftz* (Δ1–6)	130[d]
Asn 51	A-3	*bcd* Asn 51 → Ala 51	20[a]
		bcd Asn 51 → Gln 51	60[a]
		A-3 → G-3 (*ftz*)	92[c]
		A-3 → T-3	64[c]
		A-3 → C-3	25[c]
Ile 47	T-4	T-4 → A-4 (*ftz*)	14[c]
		T-4 → C-4	5.3[c]
		T-4 → G-4	4.2[c]
Gln 50	C-5′, C-6′	*ftz* Gln 50 → Lys 50	65[d]
		CC → GG	160[d]

[a] Hanes and Brent (1991). [b] Kapiloff et al. (1991). [c] Florence et al. (1991). [d] Percival-Smith et al. (1990).

Because the N-terminal arm and helix 3 make separate base-specific contacts (Figure 1B), the overall arrangement of the three helices should be critical for DNA-binding specificity. This tertiary structure should be primarily dependent on the hydrophobic core amino acids. In addition, amino acids that contact backbone phosphates are presumably critical for the proper alignment of the entire structure as a homeodomain docks with DNA. Ten positions are nearly perfectly conserved in all homeodomains (Figure 2) and half of these contribute to the hydrophobic core. Three of the 10 contact bases or backbone phosphates. Most of the remaining hydrophobic core and phosphate contact residues are highly conserved as well. These data suggest that the *Antp* and *en* structures should serve as good models for the structures and DNA contacts of nearly all homeodomains.

Binding Site Alignment

From comparisons of sequences protected in DNase I footprinting experiments [summarized in Hayashi and Scott (1990)] and selection from random oligonucleotides (Ekker et al., 1991; Florence et al., 1991), consensus and high-affinity binding site sequences have been identified for many homeodomains or homeodomain-containing proteins. The majority of homeodomains characterized thus far recognize the sequence 5′ TAAT 3′ (Odenwald et al., 1989). These include members of the *Drosophila* Antennapedia and Bithorax complex gene families as well as a number of more divergent homeodomains including *en*, even-skipped (*eve*), paired (*prd*), and bicoid (*bcd*). The portions of these binding sites corresponding to the region of base-specific contact in the *Antp* and *en* structures are aligned in Figure 3. Mutational analysis suggests that the *ftz* and Ultrabithorax (*Ubx*) homeodomains recognize base sequence at each position of the TAATNN sequence (Percival-Smith et al., 1990; Ekker et al., 1991; Florence et al., 1991). Flanking sequences are only weakly discriminated by the *ftz* homeodomain (Florence et al., 1991) but may be recognized by the full-length proteins, which in some cases have been shown to contain additional DNA-binding domains (see below).

The *Antp* and *en* homeodomain–DNA complex structures show that the first two base pairs of TAATNN are contacted in the minor groove by the N-terminal arm, while the last four base pairs are contacted in the major groove by helix 3 (Figures 1B and 4). The high-affinity binding site contained in the *en* structure actually consists of two overlapping and symmetrically arranged TAAT sequences (TAATTA), which alone would leave uncertainty in determining the alignment of binding sites sequences. However, comparison with the *Antp* structure makes it quite clear that minor-groove contact occurs at the 5′ end of the core TAAT sequence. Methylation interference and mutational analysis of the *ftz* and *bcd* homeodomains and their binding sites have independently led to the same alignment of helix 3 and the N-terminal arm with respect to the TAATNN sequence (Percival-Smith et al., 1990; Hanes & Brent, 1991). As shown in Figure 3, a similar alignment has been determined for the *oct-1* POU homeodomain (Verrijzer et al., 1990; Kristie & Sharp, 1990). The POU-specific domain ($\mathrm{POU_S}$), located near the N-terminus of the homeodomain, recognizes the sequence ATGC, apparently by contact in the minor groove adjacent to the site of N-terminal arm contact.

The *Antp* and *en* homeodomain–DNA structures are very similar, even though the contacts inferred by the positions of neighboring amino acids and bases in the *Antp* study differ somewhat from the contacts apparent in the *en* structure. These results may simply reflect differences in structural resolution, but it is conceivable that real differences exist since the sequences of both the homeodomains and the DNA differ. In addition, B DNA was used in modeling the *Antp* structure while helical twist, propeller twist, and base tilt vary along the length of the DNA in the *en* structure. Alternatively, complexes in solution (*Antp*) may differ slightly in structure from those packed into a crystal (*en*). In any case, mutational analysis of homeodomains and their binding sites has yielded additional information about the nature of homeodomain–DNA contacts. Together, the structural and mutational data provide a number of insights into the nature of base-specific homeodomain–DNA interactions.

Major-Groove Contacts

Three to four side chains of helix 3 make base-specific contacts with 3–4 bases in the major groove (Figure 1). In the *en* structure these are Ile 47, Gln 50, and Asn 51. The α carbons of these residues are in approximately the same positions in the *Antp* structure. In addition, Met 54 of *Antp* is in position to make base contact, while *en* contains an alanine at position 54 that does not extend far enough for base contact. In the discussion that follows, one strand of the 6-bp region of base contact, TAATNN, is numbered 1–6 (Figure 3), with 1′–6′ referring to complementary bases on the opposite strand. Positions 1–6 are TAATGG in the *Antp* structure and TAATTA in the *en* structure.

Only the highly divergent *pem* homeodomain contains an amino acid other than asparagine at position 51 (Table II). From these data, it seems that homeodomain structure constrains position 51 to Asn, preventing the evolution of new DNA-binding specificities by alteration of this residue. In the *en* structure, Asn 51 forms two hydrogen bonds with the adenine at position 3 (Figure 4) and is in approximately the same position in the *Antp* structure. Mutation of Asn 51 to either alanine or glutamine reduces *bcd* activity in yeast reporter assays 20-fold (Table I; Hanes & Brent, 1991) and mutations at position 3 reduce *ftz* homeodomain binding 25–92-fold (Florence et al., 1991). The conservation of Asn 51 suggests that all homeodomains recognize adenine at position 3.

Table II: Helix 3 Sequences That Vary at the Positions of Base-Specific Contact Identified in the *Antp* and *en* Homeodomains

	base contact position			
homeodomain	47	50	51	54
Antp	Ile	Gln	Asn	Met
HOX1 group				Val
en,[a] *cad*, *cdx*, *ro*, *msh*, Hox7.1, PHO2				Ala
ceh-1, NK-1				Thr
dlx				Ser
NK-2, NK-3, *msh-2*, TTF-1				Tyr
eve,[a] H2.0, *Hlx*	Val			
ceh-10, *Mix-1*, *Athb-1,2*, *zfh-2II*	Val			Ala
Isl-1	Val			Cys
mec-3, *Lin-11*, *ceh-14*, *zhf-1*	Val			Ser
ems, E5, *ceh-5*	Val			Thr
ceh-7	Val			Ile
prd, *gsb*	Val	Ser		Ala
POU family[a]	Val	Cys		Gln
bcd		Lys		Arg
otd, goosecoid	Val	Lys		Ala
MATa1	Val	Ile		
BarH1, HOX11	Thr			Thr
cut	Asn	His		
MATa2	Asn	Ser		Arg
mat-2-P	Asn	Ser		Arg
HNF1a, B	Asn	Ala		Ala
prl	Asn	Gly		Ile
Kn1, ZMH1,2	Asn	Ile		Lys
pem	Asn	Lys	Ile	Arg
b2	Leu	Ile		Arg
zfh-2I	His	Arg		Phe

[a] Denotes a group of homeodomains listed in Figure 2.

Amino acid 47 is more variable and, in the case of Ile 47, less specific in base recognition than Asn 51. In the *en* structure Ile 47 makes hydrophobic contact with thymine at position 4. The *Antp* structure indicates that hydrophobic contact also occurs between Ile 47 and the adenine at position 3. Mutations at position 4 cause up to 14-fold reductions in binding by the *ftz* homeodomain (Table I; Florence et al., 1991), indicating that base contact by Ile 47 is important for sequence recognition, although not as critical as contact by Asn 51. Valine is present at position 47 in many homeodomains, including even-skipped (*eve*), *prd*, and *oct-1* (Table II), each of which is capable of high-affinity binding to a TAAT-containing site (Figure 3; Hoey & Levine, 1988; Treisman et al., 1989; Verrijzer et al., 1990). Asparagine is present at position 47 in 10 divergent homeodomains, threonine is present three times, and leucine and histidine each occur only once. Although there is more variation than at position 51, it is apparent that homeodomains utilize a fairly limited number of amino acids for base contact at position 47.

Although position 50 is only one of several important sequence-specific contacts, compilation of base contact sequences suggests that it may be the principal source of major differences in specificity between distantly related homeodomains. Genetic and structural studies indicate that amino acid 50 contacts bases 5′ and 6′. Using a reporter gene assay in yeast, Hanes and Brent (1989, 1991) found that changing the lysine at position 50 in the *bcd* homeodomain to glutamine abolished recognition of the *bcd* binding site, TCTAATCCC, and conferred binding to an *en* site, ATTTAATTGA (*en* contains glutamine at position 50). The ability of glutamine vs lysine at position 50 to directly alter DNA-binding specificity was demonstrated in the context of the *prd* homeodomain by Treisman et al. (1989). Both studies also demonstrated that changes affecting amino acids corresponding to positions of base contact in prokaryotic HTH proteins did not affect binding-site specificity. Glutamine 50 is located near bases 5′ and 6′ in the *Antp* and *en* structures (Otting et al., 1990;

			10	20	30	40	50	60	
			N-term	----helix 1----		--helix 2--	----helix 3-4-----		Ref
			B B	H	PH	P HH HH H	H BHHB	B PBP P	
			*	* * *		*	* **	* *	
	Antp	D	RKRGRQTYTR	YQTLELEKEF	HFNRYLTRRR	RIEIAHALCL	TERQIKIWFQ	NRRMKWKKEN	1
HOX	HOX1J	H	GRKK-VP--K	V-LK---R-Y	AT-KFI-KDK	-RR-SATTN-	S---VT----	---V-E--VI	2
Group 1	HOX3G	H	GRKK-VP--K	V-LK-----Y	AASKFI-KEK	-RR-SATTN-	S---VT----	---V-E--VV	2
	HOX4I	H	GRKK-VP--K	L-LK---N-Y	AI-KFINKDK	-RR-SA-TN-	S---VT----	---V-D--IV	3
Group 2	HOX3F	H	SRKK-KP-SK	L-LA---G--	LV-EFI--Q-	-R-LSDR-N-	SDQ-V-----	-----K-RLL	2
	HOX4H	H	ARKK-KP--K	Q-IA---N--	LV-EFIN-QK	-K-LSNR-N-	SDQ-V-----	-----K-RVV	3
Group 3	HOX1I	H	TRKK-CP--K	--IR---R--	F-SV-INKEK	-LQLSRM-N-	-D--V-----	-----E--I-	2
	HOX4F	H	SRKK-CP--K	--IR---R--	F--V-INKEK	-LQLSRM-N-	-D--V-----	-----E--L-	2
Group 4	HOX1H	H	GRKK-CP--K	H---------	L--M----E-	-L--SRSVH-	-D--V-----	-----L--M-	2
	HOX4D	H	GRKK-CP--K	H---------	L--M----E-	-L--SRSVH-	-D--V-----	-----L--MS	2
Group 5	AbdB	D	VRKK-KP-SK	F---------	L--A-VSKQK	-W-L-RN-Q-	----V-----	-----N--NS	1
	HOX1G	H	TRKK-CP--K	H---------	L--M----D-	-Y-V-RL-N-	----V-----	-----M--I-	2
	HOX2E	H	SRKK-CP--K	----------	L--M----D-	-H-V-RL-N-	S---V-----	-----M--M-	2
	HOX3B	H	TRKK-CP--K	----------	L--M----D-	-Y-V-RV-N-	----V-----	-----M--M-	2
	HOX4C	H	TRKK-CP--K	----------	L--M----D-	-Y-V-RI-N-	----V-----	-----M--MS	2
	ceh-11	C	S-K-----Q-	---SV--AK-	QQSS-VSKKQ	-E-LRLQTQ-	-D--------	-----A---K	4
Group 6	HOX2D	H	-R------S-	----------	L--P----K-	---VS---G-	----V-----	----------	2
	HOX3A	H	-RS-----S-	----------	L--P----K-	---VS---G-	----V-----	----------	2
	HOX4E	H	-R------S-	F---------	L--P----K-	---VS---A-	----V-----	----------	2
Group 7	Antp	D	----------	----------	----------	----------	----------	----------	1
	Ubx	D	-R--------	----------	-T-H------	---M------	----------	-----L---I	1
	abdA	D	-R--------	F---------	---H------	----------	----------	-----L---L	1
	HOX1A	H	----------	----------	----------	----------	----------	---------H	2
	HOX2C	H	----------	----------	-Y--------	----------	----------	----------	2
	mab-5	C	S--T----S-	S---------	-YHK----K-	-Q--SET-H-	----V-----	-----H---A	1
Group 8	Scr	D	T--Q-TS---	----------	----------	----------	----------	---------H	1
	HOX1B	H	GR--------	----------	----------	-----N----	----------	----------	2
	HOX2B	H	GR--------	----------	-Y--------	----------	----------	---------S	2
	HOX3C	H	-R----I-S-	----------	----------	-----N----	----------	---------S	2
Group 9	ftz	D	S--T------	----------	-----I----	--D--N--S-	S---------	-----S--DR	1
	HOX1C	H	G--A-TA---	----------	----------	----------	S---------	--------D-	2
	HOX2A	H	G--A-TA---	----------	----------	----------	S---------	--------D-	2
	HOX3D	H	G--S-TS---	----------	----------	-----NN---	N---------	--------DS	2
Group 10	Dfd	D	P--Q-TA---	H-I-------	-Y--------	------T-V-	S---------	--------D-	1
	HOX1D	H	P--S-TA---	Q-V-------	----------	------T---	S---V-----	--------DH	2
	HOX2F	H	P--S-TA---	Q-V-------	-Y--------	-V--------	S---------	--------DH	2
	HOX3E	H	P--S-AA---	Q-V-------	-Y--------	------S---	S---------	--------DH	2
	HOX4B	H	P--S-TA---	Q-V-------	----------	------T---	S---------	--------DH	2
	ceh-15	C	E--Q-TA---	N-V-------	-THK----K-	---V--S-M-	----V-----	-----H----	5
Group 11	zen1	D	L--S-TAF-S	V-LV---N--	KS-M--Y-T-	-----QR-S-	C---V-----	-----F--DI	2
	zen 2	D	S--S-TAFSS	L-LI---R--	-L-K--A-T-	----SQR-A-	----V-----	-----L--ST	2
	HOX2G	H	S--A-TA--S	A-LV------	------C-P-	-V-M-NL-N-	S---------	-----Y--DQ	2
	HOX4A	H	S--V-TA--S	A-LV------	------C-P-	-V-M-NL-N-	-----L----	-----Y--DQ	2
Group 12	HOX2H	H	AR-L-TA--N	T-L-------	---K--C-P-	-V---AL-D-	----V-V---	-----H-RQT	2
	pS6	S	PG-L-TA--N	T-L-------	---K--C-P-	-V---AL-D-	----V-V---	-----H-RQT	2
Group 13	lab	D	NNS--TNF-N	K-LT------	--------A-	-----NT-Q-	N-T-V-----	-----Q--RV	2
	HOX2I	H	PSGL-TNF-T	R-LT------	---K--S-A-	-V---AT-E-	N-T-V-----	-----Q--RE	2
	ceh-13	C	NGTN-TNF-T	H-LT------	-TAK-VN-T-	-T---SN-K-	Q-A-V-----	-----E--RE	5
EN	en	D	E--P-TAFSS	E-LAR-KR--	NE-----E--	-QQLSSE-G-	N-A-------	-K-A-I--ST	1
	inv	D	D--P-TAFSG	T-LAR-KH--	NE-----EK-	-QQLSGE-G-	N-A-------	-K-A-L--SS	1
	en1	M	D--P-TAF-A	E-LQR-KA--	QA---I-EQ-	-QTL-QE-S-	N-S-------	-K-A-I--AT	1
	en2	M	D--P-TAF-A	E-LQR-KA--	QT-----EQ-	-QSL-QE-S-	N-S-------	-K-A-I--AT	1
	suhb-en	SU	E--P-TAFSA	S-LQR-KQ--	QQSN---EQ-	-RSL-KE-T-	S-S-------	-K-A-I--AS	1
EVE	eve	D	VR-Y-TAF--	D-LGR-----	YKEN-VS-P-	-C-L-AQ-N-	P-ST--V---	-----D-RQR	1
	Evx 1	M	MR-Y-TAF--	E-IAR-----	YREN-VS-P-	-C-L-A--N-	P-TT--V---	-----D-RQR	3
	Evx 2	M	VR-Y-TAF--	E-IAR-----	YREN-VS-P-	-C-L-A--N-	P-TT--V---	-----D-RQR	3

```
                                    10         20         30         40         50         60
                            N-term   ----helix 1----     --helix 2--     ----helix 3-4-----
                              B B           H         PH     P  HH HH H     H BHHB B PBP P
                                *       *   *   *                     *     *  **  * *
          Antp                  RKRGRQTYTR YQTLELEKEF HFNRYLTRRR RIEIAHALCL TERQIKIWFQ NRRMKWKKEN
PRD       prd         D     QR-C-T-FSA S-LD---RA- ERTQ-PDIYT -E-L-QRTN- --AR-QV--S ---ARLR-QH   1
          gsbBSH4     D     QR-S-T-F-A E-LEA--RA- SRTQ-PDVYT -E-L-QTTA- --AR-QV--S ---ARLR-HS   1
          gsbBSH9     D     QR-S-T-FSN D-IDA--RI- ARTQ-PDVYT -E-L-QSTG- --ARVQV--S ---ARLR-QL   1
          ceh-10      C     KR-H-TIF-Q --ID----A- QDSH-PDIYA -EVL-GKTE- Q-DR-QV--- ---A--R-TE   4
          otd         D     QR-E-T-F-- A-LDV--AL- GKT--PDIFM -E-V-LKIN- P-SRVQV--K ---A-CRQQL   6
          Mix-1       X     QR-K-TFF-Q A-LDI--QF- QT-M-PDIHH -E-L-RHIYI P-SR-QV--- ---A-VRRQ    4
          goosecD     X     KR-H-TIF-D E-LEA--NL- QETK-PDVGT -EQL-RRVH- R-EKVEV--K ---A--RRQK   7

          bcd         D     PR-T-T-F-S S-IA---QH- LQG----AP- LADLSAK-A- GTA-V----K ---RRH-IQS   1

          ro          D     QR-Q-T-FST E---R--V-- -R-E-IS-S- -F-L-ET-R- --T------- ---A-D-RIE   1

DLL       dll         D     MRKP-TI-SS L-LQQ-NRR- QRTQ--ALPE -A-L-AS-G- -QT-V----- -----Y--MM   8
          dlx         M     IRKP-TI-SS L-LQA-NRR- QQTQ--ALPE -A-L-AS-G- -QT-V----- -K-S-F--LM   8

MSH/NK    msh         D     NRKP-TPF-T Q-L-S---K- REKQ--SIAE -A-FSSS-R- --T-V----- ---A-A-RLQ   9
          Hox-7       M     NRKP-TPF-T A-L-A--RK- RQKQ--SIAE -A-FSSS-S- --T-V----- ---A-A-RLQ   10
          ceh-1       C     MR-A-TAF-Y E-LVRV-NK- LTS---SVVE -LNL-IQ-Q- S-T-V----- ---T----H-   4
          S59/NK-1    D     PR-A-TAF-Y E-LVS--RK- KTT---SVCE -LNL-LS-S- --T-V----- ---T----QN   9
          NK-3        D     K--S-AAFSH A-VF---RR- AQQ---SGPE -S-M-KS-R- --T-V----- ---Y-T-RKQ   9
          NK-1        D     KRKP-VLFSQ A-V----CR- RLKK---GAE -EI--QK-N- SAT-V----- ---Y-S-RGD   9
          EMS         D     P--I-TAFSP S-L-K--HA- ES-Q-VVGAE -KAL-QN-N- S-T-V-V--- ---T-H-RMQ   27
          E5          D     P--V-TAFSP T-L-K--HA- EG-H-VVGAE -KQL-QG-S- --T-V-V--- ---T-H-RMQ   27
          NK-2        D     KRKR-VLF-K A--Y---RR- RQQ---SAPE -EHL-SLIR- -PT-V----- -H-Y-T-RAQ   9
          TTF-1       R     -RKR-VLFSQ A-VY---RR- KQQK--SAPE -EHL-SMIH- -PT-V----- -H-Y-M-RQA   11
          cut         D     S-KQ-VLFSE E-KEA-RLA- ALDP-PNVGT IEFL-NE-G- AT-T-TN--H -H--RL-QQV   1

H2.0      H2.0        D     -SWS-AVFSN L-RKG--IQ- QQQK-I-KPD -RKL-AR-N- -DA-V-V--- ------RHTR   1
          Hlx/HB24    H/M   -SWS-AVFSN L-RKG---R- EIQK-V-KPD -KQL-AM-G- -DA-V-V--- ------RHSK   12
          AHox1       AS    --WN-AVFSL M-RRG---S- QSQK-VAKPE -RKL-D--S- -DA-V----- ------RQ-I   13

BAR       DmBarH1     D     QRKA-TAF-D H-LQT---S- ERQK--SVQE -Q-L--K-D- SDC-V-T-Y- ---T---RQT   14
          DaBarH1     D     QRKA-TAF-D H-LAT---S- ERQK--SVQE -Q-LS-K-D- SDC-V-T-Y- ---T---RQT   14
          HOX11       H     K-KP-TSF-- L-IC----R- -RQK--ASAE -AAL-K--KM -DA-V-T--- ---T--RRQT   15

CAD       ceh-3       C     ADKY-MV-SD --R------- -TSPFI-SD- KSQLSTM-S- ---------- --           17
          cad         D     KDKY-VV--D F-R------Y CTS--I-I-- KS-L-QT-S- S---V----- ---A-ERTS-   4
          Cdx1        M     KDKS-VV--D H-R------- -YS--I-I-- KS-L-AN-G- ----V----- ---A-ER-V-   4

LIM/ZFH   Isl-1       R     TT-V-TVLNE K-LHT-RTCY AA-PRPDALM KEQLVEMTG- SP-V-RV--- -K-C-D--RS   18
          mec-3       C     -RGP-T-IKQ N-LDV-NEM- SNTPKPSKHA -AKL-LETG- SM-V-QV--- ---S-ERRLK   18
          Lin-11      C     -RGP-T-IKA K-LET-KNA- AATPKP--HI -EQL-AETG- NM-V-QV--- ---S-ERRMK   18
          ceh-14      C                         AY QTSSKPA-HV -EQL-SETG- DM-VVQV--- ---S-ERRLK   17
          zfh-1       D     KV-V-TAINE E-QQQ-KQHY SL-ARPS-DE FRM--AR-Q- DP-VVQV--- -N-SRER-MQ   19
          zfh-2I      D     Q--A-TRI-D D-LKI-RAH- DI-NSPSEES IM-MSQKAN- PMKVV-H--R -TLF-ERQRN   19
          zfh-2II     D     KRAN-TRF-D --IKV-QEF- EN-S-PKDSD LEYLSKL-L- SP-V-VV--- -A-Q-QR-IY   19
          zfh-2III    D     N--L-T-ILP E-LNF-YECY QWEWNPW-KM LE--SKKVN- KK-VVQV--- -S-A-D--SR   19

CEH       ceh-5       C     P--P-TDNAD E-LEK--ES- NTSG--SGST -AKL-ES-G- SDN-V-V--- ---T-Q--ID   17
          ceh-7       C     IP-R-T-F-V E-LYL--MY- AQSQ-VGCDE -ERL-RI-S- D-Y-V----- ---IRMRREA   17
          ceh-9       C     --KA-T-FSG K-VF----Q- EAKK--SSSD -S-L-KR-DV --T-V----- ---T----IE   4
          ceh-14      C                         AY QTSSKPA-HV -EQL-SETG- DM-VVQV--- ---A-E-RLK   17

POU I     pit-1       R     KRKR-T-ISI AAKDA--RH- GEHSKPSSQE IMRMAEE-N- EKEVVRV--C ---QRE-RVK   19

    II    oct-1       H     -RKK-TSIET NIRVA---S- LE-QKP-SEE ITM--DQ-NM EKEV-RV--C ---Q-E-RI-   19
          oct-2       H     -RKK-TSIET NVRFA---S- LA-QKP-SEE ILL--EQ-HM EKEV-RV--C ---Q-E-RI-   19
          pdm-1       D     -RKK-TSIET TIRGA---A- LA-QKP-SEE ITQL-DR-SM EKEVVRV--C ---Q-E-RI-   20
          pdm-2       D     -RKK-TSIET TVRTT---A- LM-CKP-SEE ISQL-ER-NM DKEV-RV--C ---Q-E-RI-   20

    III   cf1A        D     KRKK-TSIEV SVKGA--QH- -KQPKPSAQE ITSL-DS-Q- EKEVVRV--C ---Q-E-RMT   19
          brn-1       R     KRKK-TSIEV SVKGA--SH- LKCPKPSSQE ITNL-DS-Q- EKEVVRV--C ---Q-E-R     19
          brn-2       R     KRKK-TSIEV SVKGA--SH- LKCPKPSAQE ITSL-DS-Q- EKEVVRV--C ---Q-E-R     19
          tst-1       R     KRKK-TSIEV GVKGA--SH- LKCPKPSAHE ITGL-DS-Q- EKEVVRV--C ---Q-E-RMT   19
          ceh-6       C     KRKK-TSIEV NVKSR--FH- QS-QKPNAQE ITQV-ME-Q- EKEVVRV--C ---Q-E-RIA   19

    IV    unc-86      C     K-RK-TSIAA PEKR---QF- KQQPRPSGE- IAS--DR-D- KKNVVRV--C -Q-Q-Q-RDF   19
          brn-3       R     K-RK-TSIAA PEKRS--AY- AVQPRPSSEK IAA--EK-D- KKNVVRV--C -Q-Q-Q-R     19
          i-pou       D     GEKK-TSIAA PEKRS--AY- AVQPRPSGEK IAA--EK-D- KKNVVRV--C -Q-Q-Q-RIV   19

    V     oct3/4      M     ---K-TSIEN RVRWS--TM- LKCPKPSLQQ ITH--NQ-G- EKDVVRV--C ---Q-G-RSS   19
```

```
                                           10         20         30         40         50         60
                                   N-term   ----helix 1----     --helix 2--     ----helix 3-4-----
                                     B B           H         PH     P  HH HH H     H BHHB B PBP P
                                       *       *   *                         *     *  **  * *
            Antp       D           RKRGRQTYTR YQTLELEKEF HFNRYLTRRR RIEIAHALCL TERQIKIWFQ NRRMKWKKEN

HNF         HNF1α      R                                      VEECNRAECIQRGVSPSQ
                                   GR-N-FKWGP ASQQI-FQAY ERQKNPSKEE -ETL-QG-GV --VRVYN--A ---KEEAFRH     21
                                                                          SNL

            HNF1β      M                                      VEECNRAECLQRGCSPSK
                                   MR-N-FKWGP ASQQI-YQAY DRQKNPSKEE -EAL--G-GV --VRVYN--A ---KEEAFRQ     21
                                                                          SNL

ATHB        Athb-1     A           LPEKKRRL-T E-VHL---S- ETENK-EPE- KTQL-KK-G- QP--VAV--- ---AR--TKQ     22
            Athb-2     A           NS-KKLRLSK D-SAI--ET- KDHST-NPKQ KQAL-KQ-G- RA--VEV--- ---ART-LKQ     22

KN          Kn1        ZM                              DQH                                               23
                                   K-KKKGKLPK EARQQ-LSWW YKWP-PSETQ KVAL-ESTG- DLK--NN--I -Q-KRHWKPS

            ZMH2       ZM                              QAH                                               23
                                   ---RAGKLPG DTAST-KAWW SKWP-P-EED KARLVQETG- QLK--NN--I -Q-KRNWHN-

            ZMH1       ZM                              QEH                                               23
                                   ---RAGKLPG DT-SI-KQWW SKWP-P-EDD KAKLVEETG- QLK--NN--I -Q-KRNWHN-

            prl        H                               YSH
                                   AR-K-RNFNK QA-EI-NEY- LSNP-PSEEA KE-L-KKCGI -VS-VSN--G -K-IRY--NI     24

            pem        M           QMPLQGSFAQ HRLR---SIL QRTNSFDVP  -EDLDRLMDA CVSRVQN--K I--AAARRNR     25

FUNGAL      MATα2      Y                                 NIE
                                   KPYRGHRF-K ENVRI--SW- AK-P--DTKG LENLMKNTS- SRI---N-VS ---R-E-TIT      1
            MATa1      Y           SPK-KSSISP QARAF--QV- RRKQS-NSKE KE-V-KKCGI -PL-VRV--I -K--RS-XXX      1
            PHO2       Y           QRPK--RAKG EA-DV-KRK- EI-PTPSLVE -KK-SDLIGM P-KNVR---- ---A-LR-KQ      1
            mat2-P     Y           VRGQCSKC-K PHLMRWLLLH YD-P-PSNSE FYDLSA-TG- -RT-LRN--S ---R            1

            b2         U                                             ESTARVGLSKANRPP
                                   CRDLSEDLPA -HMRKHFLHT LD-P-P-QEE KEGLVRLTN- EVH-LTL--I -A-RRSGWSH     26
```

FIGURE 2: Updated and abridged list of homeodomain sequences first compiled by Scott et al. (1989). Where possible, homeodomains have been assigned to subgroups; the grouping of HOX loci is according to Acampora et al. (1989). To minimize redundancy, in most subgroups the only vertebrate homeodomains listed are either human or mouse. Matches to the *Antp* sequence shown at the top of each page appear as dashes. Numbering of amino acids is at the top with amino acid 1 corresponding to amino acid 2 in Scott et al. (1989). Positions corresponding to the hydrophobic core (H), base contact (B), and phosphate contact (P) amino acids of the *Antp* and *en* structures are indicated directly above the *Antp* sequence. Positions that are invariant or nearly invariant are indicated with asterisks. Organism abbreviations are in parentheses after the name: A, *Arabidopsis*; AS, ascidian; D, *Drosophila*; C, *Caenorhabditis elegans*; H, human; M, mouse; R, rat; S, salmon; SU, sea urchin; U, *Ustilago maydis*; X, *Xenopus*; Y, yeast; Z, zebrafish; ZM, *Zea mays*. References or previous compilations containing lists of the original references for these sequences are (1) Scott et al. (1989); (2) Acampora et al. (1989); (3) D'Esposito et al. (1991); (4) Hawkins and McGhee (1990); (5) Kenyon and Wang (1991); (6) Finkelstein et al. (1990); (7) Blumberg et al. (1991); (8) Price et al. (1991); (9) Kim and Nirenberg (1989); (10) Robert et al. (1989); (11) Guazzi et al. (1990); (12) Allen et al. (1991), Deguchi et al. (1991); (13) Saiga et al. (1991); (14) Kojima et al. (1991); (15) Hotano et al. (1991); (16) Burglin et al. (1989); (17) Freyd et al. (1990), Karlsson et al. (1990); (18) Fortini et al. (1991); (19) Rosenfeld (1991); (20) Billin et al. (1991); (21) Mendel et al. (1991), De Simone et al. (1991); (22) Ruberti et al. (1991); (23) Vollbrecht et al. (1991); (24) Nourse et al. (1990), Kamps et al. (1990); (25) Rayle (1991), Sasaki et al. (1991); (26) Schulz et al. (1990); (27) Dalton et al. (1989).

Kissinger et al., 1990). In the *en* structure, Gln 50 makes hydrophobic contact with the thymine at position 6′ (Figure 4) and modest changes in DNA conformation would allow hydrogen bonding to the adenine at position 5′. As shown in Table III, the *ftz* homeodomain is capable of recognizing several dinucleotide combinations at base pairs 5 and 6. These data suggest that, in the context of *ftz*, both Gln 50 and Lys 50 are able to make specific contacts at either the 5′ or 6′ base. Although no structural data are yet available for a Lys 50 homeodomain–DNA complex, a reasonable extrapolation of the Gln 50 model supposes that Lys 50 is capable of hydrogen bonding with the proton acceptors of guanine at position 5′ or 6′. The aliphatic portion of the Lys 50 side chain may also be capable of making specific hydrophobic contacts.

Even though it recognizes TAAT, the *bcd* homeodomain appears to represent a context for base contact by amino acid 50 that differs somewhat from *ftz*. As shown in Table III, Hanes and Brent (1991) found that, unlike *ftz*, *bcd* recognizes certain sequences primarily at base pair 5 rather than at 5 and 6 (however, base pair 6 does clearly play a role in some sequence combinations). This may be because *bcd* and *ftz* differ at an additional site of potential base contact: the amino acid at position 54 is methionine in *ftz* and arginine in *bcd*. As with positions 47, 50, and 51, position 54 tends to be conserved within related groups of homeodomains (Figure 2), an indication that this position is important for homeodomain function. However, much more variation occurs at position 54 between groups than at the other positions of base contact

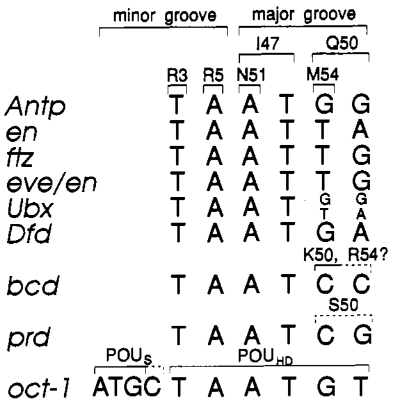

FIGURE 3: Alignment of homeodomain binding-site sequences. The positions of base contact identified in the *en*- and *Antp*-DNA complexes (Kissinger et al., 1990; Otting et al., 1990) are shown as brackets at the top. Position 50 contacts deduced for *bcd* (Hanes and Brent, 1991) and *prd* (Treisman et al. 1989) are indicated above their binding-site sequences. The *oct-1* alignment and sequence is according to Kristie and Sharp (1990). POUs is POU-specific domain; POUhd is POU homeodomain. Binding-site sequences are from the following sources: *Antp*, Muller et al. (1988), Otting et al. (1990); *en*, Kissinger et al. (1990); *eve/en*, Hoey and Levine (1988), Desplan et al. (1988); *ftz*, Pick et al. (1990); Florence et al. (1991); *Ubx*, Ekker et al. (1991); *Dfd*, Regulski et al. (1991); *bcd*, Driever and Nusslein-Volhard (1989); *prd*, Treisman et al. (1991b). Base pairs are numbered at the bottom as described in the text.

Table III: Recognition of Positions 5 and 6 by Glutamine or Lysine at Position 50 in the *ftz* and *bcd* Homeodomains

binding site[a]	*ftz* (Q50)	*ftz* (K50)	*bcd* (Q50)[b]	*bcd* (K50)[b]
T G	+[c]		+	–
G G	+[d]		–	–
C G	+[c,d]	+[d]	–	–
A G	+[c]		–	–
T A	+[c]		+	–
G A				
C A	–[c]		±	±
A A	–[c]		–	–
T C	+[c]		–	–
G C	+[d]	+[d]		
C C	–[d]	+[d]	–	+
A C	–[c]		–	–
A T	±[c]		±	–
G T				
C T			–	±
A T				

[a] Nucleotides listed occupy positions 5′ and 6′; TAAT occupies positions 1′–4′. [b] Hanes and Brent (1991). [c] Florence et al. (1991). [d] Percival-Smith et al. (1990).

(Table II), providing a potential source of extensive variation in sequence specificity. A second possibility is that overall structural differences between *ftz* and *bcd* lead to differences in how helix 3 docks with the major groove. A difference that stands out between the two homeodomains is at position 31, where *bcd* has leucine while *ftz* and most other homeodomains have arginine. Arg 31 contacts a backbone phosphate in the *en* structure and presumably in the case of *ftz* as well.

The ability of Gln or Lys at position 50 to recognize multiple dinucleotide combinations suggests that side chains at position 50 are inherently more flexible in making base contact than are the side chains of Asn 51 or Ile 47. Flexibility would also be consistent with the occurrence of a variety of amino acids at position 50 in various homeodomains (Table II). Of these, *prd* (Ser 50) and *oct-1* (Cys 50) are also able to recognize TAAT-containing sites (Figure 3), evidence that, for these homeodomains, docking of helix 3 with the major groove is similar to that of *en* and *Antp*.

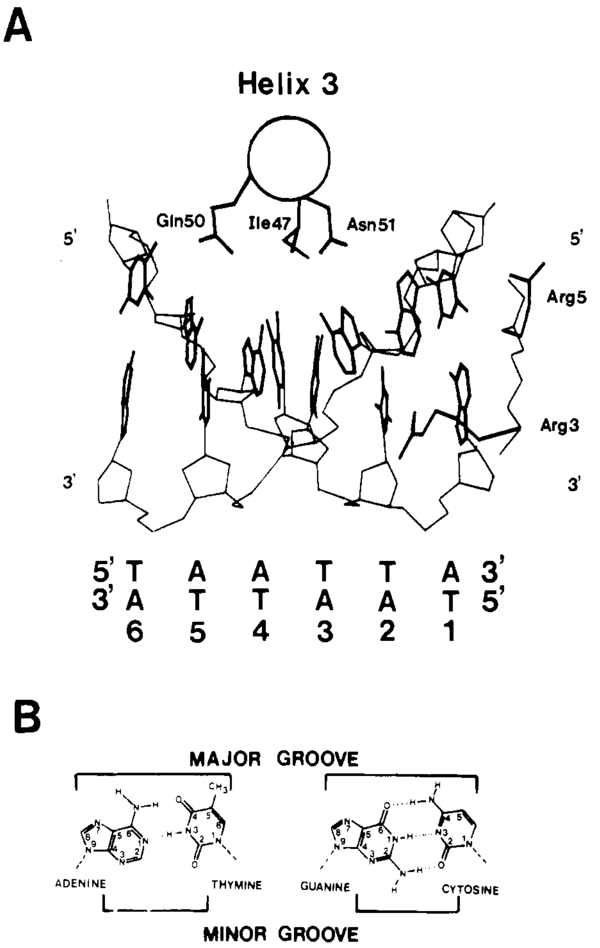

FIGURE 4: Base-specific contacts of the *en* homeodomain. At the top is a diagram, using coordinates kindly provided by C. Kissinger and C. Pabo, showing helix 3 and the N-terminal arm of *en* contacting the sequence TAATTA. A-T and G-C base pairs appear below, showing the positions of functional groups exposed to the major and minor grooves.

It will be interesting to learn whether other amino acids are either less or more selective than glutamine at position 50. It is intuitive to expect that short side chains limited to contact with only one base might result in less specific, lower affinity binding compared to side chains that contact two bases. However, since the *ftz* homeodomain can effectively recognize at least 6 out of 16 possible dinucleotide combinations at positions 5 and 6 (Table III), amino acids at position 50 that contact one rather than two bases may have a more profound effect on overall binding affinity rather than on the ability to discriminate binding-site sequence.

A number of Gln 50-containing homeodomains or full-length homeodomain proteins have been reported to have consensus sequences that differ at base pairs 5 or 6 (Figure 3). For the Ultrabithorax (*Ubx*) and *ftz* homeodomains, optimal binding-site sequences that differ only at base pairs 5 or 6 have been selected from pools of random oligonucleotides, and the absolute rankings of four position 5–6 dinucleotide combinations measured in both studies were only slightly different (Ekker et al., 1991; Florence et al., 1991). It is not known whether there are biologically significant differences in the recognition of positions 5′ and 6′ by the closely related but functionally distinct *Antp*, *Dfd*, *ftz*, and *Ubx* homeodomains.

Minor-Groove Contacts

It has long been realized that the minor groove presents fewer combinations of functional groups than the major groove and therefore contains less information than the major groove for base-pair recognition by proteins (Dickerson 1982). In spite of this, homeodomains do recognize specific bases, even A vs T, by contact of the N-terminal arm with the minor groove. Sequence conservation suggests that minor-groove binding is a universal feature of homeodomains but little is known about how N-terminal arm sequence affects minor-groove recognition.

Deletion of the N-terminal arm reduces *ftz* homeodomain binding affinity 130-fold (Percival-Smith et al., 1990). In the *en* structure, Arg 3 and Arg 5 of the N-terminal arm are in position to hydrogen bond with the O2 proton acceptors on the minor-groove sides of the thymines at positions 2′ and 1, respectively (Figure 4). Mutation of T-2′ to A-2′ reduces *ftz* homeodomain binding affinity by only 3.4-fold while C-2′ or G-2′ reduce affinity 20-fold (Table I; Florence et al., 1991). N2 of guanine projects into the center of the minor groove and thus may interfere with hydrogen bonding of Arg 3 to the O2 of cytosine in G-C base pairs (Figure 4). Arg 5 is nearly invariant in homeodomains, and lysine, which sometimes occurs instead of Arg 3, should also be limited to contact with a protein acceptor in the minor groove. This suggests that A-T base pairs at positions 1 and 2 may be optimal for recognition by nearly all homeodomains. Mutation of T-1 to either A or G reduces *ftz* binding affinity about 8-fold (Table I); apparently Arg 5 is less flexible than Arg 3 and therefore less able to make optimal contact with O2 on the wrong side of the minor groove.

Except for Arg 5, there is considerable variation in N-terminal arm sequence between, but not within, related groups of homeodomains. The significance of this variation for DNA binding affinity or specificity is unknown and difficult to predict since the N-terminus lacks regular secondary structure. A recent report has shown that the N-terminal arm of the *pit-1* homeodomain mediates posttranslational regulation of DNA binding (Kapiloff et al., 1991). *pit-1* is phosphorylated at Thr-7 in response to induction of pituitary cells with cAMP or the phorbol ester TPA. Phosphorylation at the same residue in vitro reduces the binding affinity of *pit-1* 10–20-fold for a subset of binding sites. This effect could be due to electrostatic repulsion between the phosphate on Thr 7 and the nearby backbone phosphate between bases 1 and 2, a site of contact for Thr 6 in the *en* structure (Figure 1A). However, neither Thr-6 nor Thr-7 makes any important phosphate contacts in nonphosphorylated *pit-1* since mutations changing either or both of these positions to alanine have no effect on the ability of nonphosphorylated protein to bind DNA.

Sequence Recognition by Divergent Homeodomains

Can the *Antp*/*en* model for base-specific recognition be extended to include highly divergent homeodomains? Studies of the divergent (relative to *Antp* and *en*) *pit-1* and *oct-1* POU homeodomains have shown that they contact 5–6 bp containing 5′ TAAT 3′ or a related A-T rich sequence, preceded by 5′ ATGC 3′, which is contacted by the adjacent POU$_s$ (Figure 3; Ingraham et al., 1991; Verrijzer et al., 1990; Kristie & Sharp, 1991). Although the details of base contact remain to be determined, the *oct-1* homeodomain appears to be related in specificity to *Antp* class homeodomains (Verrijzer et al. 1990).

The *Antp*/*en* model, however, apparently does not apply to the mammalian TTF-1 (thyroid transcription factor) homeodomain (Guazzi et al., 1990; Damante & Di Lauro, 1991). Even though TTF-1 and *Antp* are identical at four of six *Antp* base contact positions (5, 47, 50, and 51), the CACTCAAG consensus binding site of TTF-1 bears no obvious relationship to the *Antp* consensus, and the two proteins are unable to recognize each other's binding sites. In addition, Gln 50 is not critical for TTF-1 binding, suggesting that helix 3 of TTF-1 does not contact DNA according to the *Antp*/*en* model. The specificity differences between the two homeodomains map to sequences outside of helix 3, in both the N-terminal halves and the C-terminal ends. The structural basis for these specificity differences is unknown; however, from analysis of chimeric proteins, it appears that either the N-terminal arm or helix 1 of TTF-1 is incompatible with helix 2 of *Antp*.

Alternative Modes of DNA Sequence Recognition

Two groups have reported that individual homeodomains are capable of binding to two classes of sites. The *eve* homeodomain recognizes the TAATTG sequence but also an apparently unrelated sequence, TCAGCACCG (Hoey & Levine, 1988; Hoey et al., 1988). Treisman et al. (1989, 1991a) have found a similar phenomenon for the *prd* homeodomain, which is able to recognize both TAATCG and TTTGACGT. For *prd*, recognition of the first but not the second class of sites is dependent upon the amino acid at position 50, while recognition of the second class, but not the first, is dependent upon the amino acid at position 43 (Treisman et al., 1991b). Amino acid 43 corresponds to one of the positions of base contact in prokaryotic HTH proteins, suggesting that *prd* and perhaps additional homeodomains are capable of utilizing either of two completely different sets of DNA contact residues. The additional positions corresponding to prokaryotic HTH base contact amino acids are 42, 44, 46, and 47. These positions tend to be conserved within related homeodomain groups (Figure 2). However, Arg 43 plays no essential role in the *ftz* homeodomain, since a mutation replacing it with alanine does not prevent *ftz* from functioning normally in *Drosophila* embryos (Percival-Smith et al., 1990). It remains to be seen whether mutations altering positions 42, 43, 44, or 46 will affect the function of *prd* in *Drosophila*.

Protein–Protein Interactions

Homeodomain monomers have only about 100-fold higher affinity for specific binding sites than for nonspecific DNA, suggesting that regulatory specificity must require additional contributions to specificity. Prokaryotic HTH proteins increase their DNA-binding specificities dramatically by forming dimers (Ptashne, 1986). This also appears to be the case for at least some homeodomain proteins. The yeast MATα2 homeodomain protein represses transcription by binding to DNA as either a MATα2–MATα2 or a MATα2–MATa1 dimer, with protein–protein contact occurring through a region of the protein separate from the homeodomain (Goutte & Johnson, 1988; Sauer et al., 1988). The homeodomain-containing mammalian tissue-specific activators HNF1α and HNF1β form homo- and heterodimers via a myosin-like dimerization helix located at a distance from the homeodomain (Nicosia et al., 1990; De Simone et al., 1991; Mendel et al., 1991). Two plant homeodomain proteins, *Athb-1* and *Athb-2*, contain leucine zippers as well, suggesting that they also form dimers (Ruberti et al. 1991). The mammalian *pit-1* protein exists as a monomer in solution but binds to DNA cooperatively alone or with the *oct-1* protein (Ingraham et al. 1990; Voss et al. 1991). *pit–pit* and *pit–oct* cooperativity is mediated by the POU-specific domains and not the POU homeodomains of these proteins. In each case where dimerization or cooperative binding to DNA occurs, it is mediated by sequences outside of the homeodomain.

The *Drosophila Antp*, *en*, and fushi tarazu (*ftz*) homeodomains bind to DNA as monomers in vitro (Affolter et al., 1989; Kissinger et al., 1990; Florence et al., 1991) and the presence of only nonsymmetrical homeodomain binding sites in *bcd-*, *ftz-*, *eve-*, and *Dfd*-regulated enhancers suggests that these proteins bind to DNA as monomers in vivo (Driever & Nusslein-Volhard, 1989; Kassis et al., 1989; Kassis 1990; Pick et al., 1990; Jiang et al., 1991; Regulski et al., 1991). In light of the low specificity of homeodomain monomers and the lack of additional DNA-binding domains in these *Drosophila* homeodomain proteins, how is specific activation of these enhancers achieved? The view emerging for eukaryotic transcriptional regulation in general is that a high degree of regulatory specificity is achieved through the combined action of multiple, relatively low affinity interactions between proteins bound to nearby DNA sequences (Frankel & Kim, 1991). Evidence for such a model is found in the specific autoregulation of the *eve* promoter, which requires the action of two additional nuclear factors that recognize sites adjacent to two *eve* binding sites (Jiang et al. 1991). Mutation of *eve* binding sites or the sites for either of the two other factors in oligomerized versions of the *eve* enhancer drastically reduces autoactivation in vivo, evidence that the three components interact synergistically (although some of this effect is undoubtedly due to the use of repeated copies of the enhancer).

The best understood interactions between homeodomain proteins and other factors involve the MATα2 protein of yeast and the mammalian *oct-1* protein. In α cells, MATα2 binds cooperatively with the MCM1 protein to sequences upstream of a-specific genes, repressing their transcription (Keleher et al., 1988). In a/α diploids, MATα2 binds with the MATa1 protein to sequences upstream of haploid-specific genes, repressing their transcription (Goutte & Johnson, 1988). In mammalian cells the *oct-1* protein promotes HSV enhancer activity by binding cooperatively with the viral VP16 protein (Stern et al., 1989) and an additional cellular factor (Kristie & Sharp, 1990). The closely related *oct-2* protein does not interact with VP16, and helix swap experiments were used to localize the site of VP16 contact on *oct-1* to helix 2 (Stern et al., 1989). Protein–protein interactions must place additional constraints on the evolution of homeodomains and may help to account for the conservation of outward-facing amino acids within homeodomain groups (Figure 2).

Protein–protein interactions may also help account for the well-documented differences in regulatory specificity between closely-related proteins of the *Drosophila* homeotic genes. This is an important issue since both these genes and their mammalian homologues play central roles in development. The *Ubx*, *Dfd*, *Antp*, and *Sex combs reduced* (*Scr*) homeodomains have identical base contact residues and at least *Antp* and *Ubx* have very similar DNA binding specificities in vitro (Affolter et al., 1990; Ekker et al., 1991), yet each protein generates a different phenotype when ectopically expressed in *Drosophila* embryos or larvae (Kuziora & McGinnis, 1989; Gibson et al., 1990; Gehring, 1990; Mann & Hogness, 1990). By swapping of sequences between *Ubx* and *Dfd*, *Ubx* and *Antp*, and *Antp* and *Scr*, the specificity differences of these proteins have been localized to the homeodomains or sequences immediately surrounding them (the proximity of the N-terminal arm and of helix 3 to nonconserved sequences that flank homeodomains suggests that flanking sequences might easily alter DNA-binding specificity). From these results, it appears that homeodomain DNA-binding specificity alone is unlikely to account completely for the differences in regulatory specificity between these proteins.

Alternatively, minor differences in DNA-binding specificity may contribute to differential transcriptional regulation by different homeotic proteins. This idea is supported by the fact that *Drosophila* is sensitive to the dosage of several homeotic genes (animals heterozygous for *Ubx*, *Scr*, or *Abd-B* exhibit homeotic transformations of segmental identity), suggesting that relatively small differences in the amount of protein bound to DNA (or available to bind) could tip the balance toward either activation or repression of transcription for a subset of regulated genes.

A number of homeodomain proteins achieve additional specificity by the coupling of homeodomains to additional DNA-binding domains (Hayashi & Scott, 1990). Associated DNA-binding domains include the POU-specific domain (Ingraham et al., 1990; Verrijzer et al., 1990), the paired domain (Treisman et al., 1991a), the LIM domain [a cysteine-rich domain which may bind to DNA (Freyd et al., 1990; Karlsson et al., 1990)], C_2H_2 zinc fingers, and even other homeodomains (Fortini et al., 1991). The *Drosophila zhfII* protein contains an astonishing three homeodomains and 16 zinc fingers. This leads to an interesting prediction: if the *zhfII* homeodomains have dissociation kinetics similar to those of the *Antp* and *ftz* homeodomains (Affolter et al., 1990; Florence et al., 1991), complexes in which two or three linked homeodomains are bound to DNA should be stable for as long as the *zhfII* protein remains active (Florence et al. 1991).

Conclusions

Although the *Antp*/*en* model for homeodomain–DNA interactions is quite detailed, it fails to explain the different binding modes of the TTF-1 and *prd* homeodomains. With regard to the *Antp*/*en* model for DNA contact, it appears from base specificity and from the conservation of base contact amino acids among homeodomains that differences in DNA binding specificity are likely to arise mainly from differences in sequence at position 50 and possibly position 54. The tighter constraints on positions 47 and 51 would be consistent with a fairly restricted orientation of helix 3 as it docks with the major groove, probably due to numerous backbone phosphate contacts. These constraints on homeodomain–DNA contact may make it possible to determine some rules for base-pair recognition by groups of homeodomains with related amino acids at position 50 and 54. Initial hope that a "recognition code" would emerge from structural studies of HTH proteins subsided as it become apparent that different HTH recognition helices dock with the major groove differently (Harrison & Aggarwal, 1990). It is obviously important to learn more about the nature of homeodomain binding mechanisms that differ from the *Antp*/*en* paradigm before the prospects of deciphering a widely applicable code for homeodomain–DNA interactions can be reasonably assessed.

Finally, for the most part, homeodomains proteins are transcription factors in need of regulatory targets. There are only a handful of genes known to be regulated by any of the *Drosophila* homeodomain proteins, although the situation is better for mammalian tissue-specific homeodomain (Rosenfeld, 1991). Since it is a virtual certainty that most homeodomain proteins act in combinatorial fashion with other transcription factors (Han et al., 1989; Vincent et al., 1989; Stern et al., 1989; Jiang et al., 1991), more extensive characterization of enhancer or promoter sequences which are bound by homeodomains and other factors will be essential for understanding the normal role of DNA-binding specificity in the regulation of transcription.

Acknowledgments

I am grateful to Phil Anderson, Michael Boedigheimer, Brian Florence, Heidi Nelson, and Duane Price for their helpful comments on the manuscript and to Claude Desplan for communicating unpublished results.

References

Acampora, D., D'Esposito, M., Faiella, A., Pannese, M., Migliaccio, E., Morelli, F., Stornaiuolo, A., Nigro, V., Simeone, A., & Bonicelli, E. (1989) *Nucleic Acids Res. 17*, 10385–10403.

Affolter, M., Percival-Smith, A., Muller, M., Leupin, W., & Gehring, W. J. (1990) *Proc. Natl. Acad. Sci. U.S.A. 87*, 4093–4097.

Affolter, M., Percival-Smith, A., Muller, M., Billeter, M., Qian, , Y. Q., Otting, G., Wuthrich, K., & Gehring, W. J. (1991) *Cell 64*, 879–880.

Aggarwal, A. K., Rodgers, D. W., Drottar, M., Ptashne, M., & Harrison, S. C. (1988) *Science 242*, 899–902.

Akam, M. (1987) *Development 101*, 1–22.

Allen, J. D., Lints, T., Jenkins, N. A., Copeland, N. G., Strasser, A., Harvey, R. P., & Adams, J. M. (1991) *Genes Dev. 5*, 509–520.

Billin, A. N., Cockerill, K. A., & Poole, S. J. (1991) *Mech. Dev. 34*, 75–84.

Blumberg, B., Wright, V. E., De Robertis, E. M., & Cho, K. W. Y. (1991) *Science 253*, 194–196.

Burglin, T. R., Finney, M., Coulson, A., & Ruvkun, G. (1989) *Nature 341*, 239–243.

Dalton, D., Chadwick, R., & McGinnis, W. (1989) *Genes Dev. 3*, 1940–1956.

Damante, G., & Di Lauro, R. (1991) *Proc. Natl. Acad. Sci. U.S.A. 88*, 5388–5392.

D'Esposito, M., Morelli, F., Acampora, D., Migliaccio, E., Simeone, A., & Boncinelli, B. (1991) *Genomics 10*, 43–50.

Deguchi, Y., Moroney, J. F., Wilson, G. L., Fox, C. H., Winter, H. S., & Kehrl, J. H. (1991) *New Biol. 3*, 353–363.

De Simone, V., De Magistris, L., Lazzaro, D., Gerstner, J., Monaci, P., Nicosia, A., & Cortese, R. (1991) *EMBO J. 10*, 1435–1443.

Desplan, C., Theis, J., & O'Farrell, P. H. (1985) *Nature 318*, 630–635.

Desplan, C., Theis, J., & O'Farrell, P. H. (1988) *Cell 54*, 1081–1090.

Dickerson, R. E. (1983) *Sci. Am. 249*, 94–111.

Driever, W., & Nusslein-Volhard, C. (1989) *Nature 337*, 138–143.

Ekker, S. C., Young, K. E., von Kessler, D. P., & Beachy, P. A. (1991) *EMBO J. 10*, 1179–1186.

Finkelstein, R., Smouse, D., Capaci, T. M., Spradling, A. C., & Perrimon, N. (1990) *Genes Dev. 4*, 1516–1527.

Florence, B., Handrow, R., & Laughon, A. (1991) *Mol. Cell. Biol. 11*, 3613–3623.

Fortini, M. E., Lai, Z., & Rubin, G. M. (1991) *Mech. Dev. 34*, 113–122.

Frankel, A. D., & Kim, P. S. (1991) *Cell 65*, 717–719.

Freyd, G., Kim, S. K., & Horvitz, H. R. (1990) *Nature 344*, 876–879.

Gibson, G., Schier, A., LeMotte, P. K., & Gehring, W. J. (1990) *Cell 62*, 1087–1103.

Goutte, C., & Johnson, A. D. (1988) *Cell 52*, 875–882.

Guazzi, S., Price, M., De Felice, M., Damante, G., Mattei, M., & Di Lauro, R. (1990) *EMBO J. 9*, 3631–3639.

Han, K., Levine, M. S., & Manley, J. L. (1989) *Cell 56*, 573–583.

Hanes, S. D., & Brent, R. (1989) *Cell 57*, 1275–1283.

Hanes, S. D., & Brent, R. (1991) *Science 251*, 426–430.

Harrison, S. C., & Aggarwal, A. K. (1990) *Annu. Rev. Biochem. 59*, 933–969.

Hatano, M., Roberts, C. W. M., Minden, M., Crist, W., & Korsmeyer, S. J. (1991) *Science 253*, 79–82.

Hawkins, N. C., & McGhee, J. D. (1990) *Nucleic Acids Res. 18*, 6101–6106.

Hayashi, S., & Scott, M. P. (1990) *Cell 63*, 883–894.

Hoey, T., & Levine, M. (1988) *Nature 332*, 858–861.

Hoey, T., Warrior, R., Manak, J., & Levine, M. (1988) *Mol. Cell. Biol. 8*, 4598–4607.

Ingraham, H. A., Flynn, S. E., Voss, J. W., Albert, V. R., Kapiloff, M. S., Wilson, L., & Rosenfeld, M. G. (1990) *Cell 61*, 1021–1033.

Jiang, J., Hoey, T., & Levine, M. (1991) *Genes Dev. 5*, 265–277.

Jordan, S. R., & Pabo, C. O. (1988) *Science 242*, 893–899.

Kamps, M. P., Murre, C., Sun, X., & Baltimore, D. (1990) *Cell 60*, 547–555.

Kapiloff, M. S., Farkash, Y., Wegner, M., & Rosenfeld, M. G. (1991) *Science 253*, 786–789.

Karlsson, O., Thor, S., Norberg, T., Ohlsson, H., & Edlund, T. (1990) *Nature 344*, 879–882.

Kassis, J. A. (1990) *Genes Dev. 4*, 433–443.

Kassis, J. A., Desplan, C., Wright, D. K., & O'Farrell, P. H. (1989) *Mol. Cell. Biol. 9*, 4304–4311.

Keleher, C. A., Goutte, C., & Johnson, A. D. (1988) *Cell 53*, 927–936.

Kenyon, C., & Wang, B. (1991) *Science 253*, 516–517.

Kessel, M., & Gruss, P. (1990) *Science 249*, 374–379.

Kim. Y., & Nirenberg, M. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 7716–7720.

Kissinger, C. R., Liu, B., Martin-Blanco, E., Kornberg, T. B., & Pabo, C. O. (1990) *Cell 63*, 579–590.

Kojima, T., Ishimaru, S., Higashijima, S., Takayama, E., Akimaru, H., Sone, M., Emori, Y., & Saigo, K. (1991) *Proc. Natl. Acad. Sci. U.S.A. 88*, 4343–4347.

Kristie, T. M., & Sharp, P. A. (1990) *Genes Dev. 4*, 2383–2396.

Kuziora, M. A., & McGinnis, W. (1989) *Cell 59*, 563–571.

Laughon, A., & Scott, M. P. (1984) *Nature 310*, 25–31.

Maier, D., Preiss, A., & Powell, J. R. (1990) *EMBO J. 9*, 3957–3966.

Mann, R. S., & Hogness, D. S. (1990) *Cell 60*, 597–610.

McGinnis, W., Garber, R. L., Wirz, J., Kuroiwa, A., & Gehring, W. J. (1984) *Cell 37*, 403–408.

Mendel, D. B., Hansen, L. P., Graves, M. K., Conley, P. B., & Crabtree, G. R. (1991) *Genes Dev. 5*, 1042–1056.

Muller, M., Affolter, M., Leupin, W., Otting, G., Wuthrich, K., & Gehring, W. J. (1988) *EMBO J. 7*, 4299–4304.

Nicosia, A., Monaci, P., Tomei, L., De Francesco, R., Nuzzo, M., Stunnenberg, H., & Cortese, R. (1990) *Cell 61*, 1225–1236.

Nourse, J., Mellentin, J. D., Galili, N., Wilkinson, J., Stanbridge, E., Smith, S. D., & Cleary, M. L. (1990) *Cell 60*, 535–545.

Odenwald, W. F., Garbern, J., Arnheiter, H., Tournier-Lasserver, E., & Lazzarini, R. A. (1989) *Genes Dev.* 3, 158–172.

Otting, G., Qian, Y. Q., Billeter, M., Muller, M., Affolter, M., Gehring, W. J., & Wuthrich, K. (1990) *EMBO J. 9*, 3085–3092.

Pabo, C. O., & Sauer, R. T. (1984) *Annu. Rev. Biochem. 53*, 293–321.

Percival-Smith, A., Muller, M., Affolter, M., & Gehring, W. J. (1990) *EMBO J. 9*, 3967–3974.

Pick, L., Schier, A., Affolter, M., Schmidt-Glenewinkel, T., & Gehring, W. J. (1990) *Genes Dev. 4*, 1224–1239.

Price, M., Lemaistre, M., Pischetola, M., Di Lauro, R., & Duboule, D. (1991) *Nature 351*, 748-751.

Qian, Y. Q., Billeter, M., Otting, G., Muller, M., Gehring, W. J., & Wuthrich, K. (1989) *Cell 59*, 573-580.

Rayle, R. (1991) *Dev. Biol. 146*, 255-257.

Regulski, M., Dessain, S., McGinnis, N., & McGinnis, W. (1991) *Genes Dev. 5*, 278-286.

Robert, B., Sassoon, D., Jacq, B., Gehring, W. J., & Buckingham, M. (1989) *EMBO J. 8*, 91-100.

Rosenfeld, M. G. (1991) *Genes Dev. 5*, 897-907.

Ruberti, I., Sessa, G., Lucchetti, S., & Morelli, G. (1991) *EMBO J. 10*, 1787-1791.

Saiga, H., Mizokami, A., Makabe, K. W., Satoh, N., & Mita, T. (1991) *Development 111*, 821-828.

Sasaki, A. W., Doskow, J., MacLeod, C. L., Rogers, M. B., Gudas, L. J., & Wilkinson, M. F. (1991) *Mech. Dev. 34*, 155-164.

Sauer, R. T., Smith, D. L., & Johnson, A. D. (1988) *Genes Dev. 2*, 807-816.

Scott, M. P., & Weiner, A. J. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 4115-4119.

Scott, M. P., Tamkun, J. W., & Hartzell, G. W., III (1989) *Biochim. Biophys. Acta 989*, 25-48.

Schulz, B., Banuett, F., Dahl, M., Schlesinger, R., Schafer, W., Martin, T., Herskowitz, I., & Kahmann, R. (1990) *Cell 60*, 295-306.

Shepherd, J. C., McGinnis, W., Carrasco, A. E., DeRobertis, E. M., & Gehring, W. J. (1984) *Nature 310*, 70-71.

Stern, S. A., Tanaka, M., & Herr, W. (1989) *Nature 341*, 624-630.

Treisman, J., Conczy, P., Vashishtha, M., Harris, E., & Desplan, C. (1989) *Cell 59*, 553-562.

Treisman, J., Harris, E., & Desplan, C. (1991a) *Genes Dev. 5*, 594-604.

Treisman, J., Harris, E., Wilson, D., & Desplan, C. (1991b) *Bioessays* (in press).

Verrijzer, C. P., Kal, A. J., & van der Vliet, P. C. (1990) *Genes Dev. 4*, 1964-1974.

Vincent, J.-P., Kassis, J. A., & O'Farrell, P. H. (1990) *EMBO J. 9*, 2573-2578.

Volhlbrecht, E., Veit, B., Sinha, N., & Hake, S. (1991) *Nature 350*, 241-243.

Voss, J. W., Wilson, L., & Rosenfeld, M. G. (1990) *Genes Dev. 5*, 1309-1320.

Accelerated Publications

Design, Chemical Synthesis, and Expression of Genes for the Three Human Color Vision Pigments[†]

Daniel D. Oprian,[*] Ana B. Asenjo,[‡] Ning Lee,[‡] and Sandra L. Pelletier

Graduate Department of Biochemistry, Brandeis University, Waltham, Massachusetts 02254

Received August 30, 1991; Revised Manuscript Received October 11, 1991

ABSTRACT: Color vision in humans is mediated by three pigments from retinal cone photoreceptor cells: blue, green, and red. We have designed and chemically synthesized genes for each of these three pigments. The genes were expressed in COS cells, reconstituted with 11-*cis*-retinal chromophore, and purified to homogeneity using an immunoaffinity procedure. To facilitate the immunoaffinity purification, each pigment was modified at the carboxy terminus to contain an additional eight amino acid epitope for a monoclonal antibody previously used to purify bovine rhodopsin. The spectra for the isolated pigments had maxima of 424, 530, and 560 nm, respectively, for the blue, green, and red pigments. These maxima are in excellent agreement with the maxima previously observed by microspectrophotometry of individual human cone cells. The spectra are the first to be obtained from isolated human color vision pigments. They confirm the original identification of the three color vision genes, which was based on genetic evidence [Nathans, J., Thomas, D., & Hogness, D. S. (1986) *Science 232*, 193].

Human color vision is mediated by three visual pigments present in retinal cone photoreceptor cells [cf. Boynton (1979)]. The spectra for these pigments have been determined by microspectrophotometry of individual human cone cells (Dartnall et al., 1983). The spectra show absorption maxima at 420, 530, and 560 nm for the blue, green, and red cone pigments, respectively. Despite their very different absorption maxima, these pigments all contain an identical 11-*cis*-retinal chromophore covalently attached to the protein by means of a Schiff base linkage to the ϵ-amino group of a conserved lysine residue. Thus, the different absorption maxima for the pigments arise from differences in the amino acid sequence of the individual proteins and in the interaction of these amino acids with the chromophore.

Several years ago, Nathans et al. (1986a,b) cloned the genes for the blue, green, and red pigments using a homologous rhodopsin probe. The genes were identified by the following genetic criteria.

Blue Gene. Only one gene was localized to an autosome, chromosome 7. This gene was concluded to be that of the blue

[†] This work was supported by NIH Grants EYO7965 and BRSG SO7 RR07044. S.L.P. was supported by NIH Training Grant 5T32 GM07596-11 awarded to the Graduate Department of Biochemistry at Brandeis University.

[*] To whom correspondence should be addressed.

[‡] Biology Graduate Program, Brandeis University.

0006-2960/91/0430-11367$02.50/0 © 1991 American Chemical Society