Biochimica et Biophysica Acta, 989 (1989) 25–48
Elsevier

BBACAN 87200

The structure and function of the homeodomain

Matthew P. Scott, John W. Tamkun and George W. Hartzell, III

Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO (U.S.A.)

(Received 18 October 1989)

Contents

I. Introduction

'Homeotic' transformations are developmental anomalies in which one part of the body develops in the likeness of another. Many cases of homeosis were described and the phenomenon was named by Bateson in [1]. He pointed out the possible significance of such abnormalities for developmental biology and emphasized that such altered pathways of development are observed in many organisms including plants, insects, echinoderms, crustaceans, fish, reptiles and mammals. Possible cases of homeosis in man have been observed as well (reviewed in Ref. 2). Genetically induced transformations have been powerful ways to identify genes that control growth and pattern formation during development. Homeosis has been more extensively studied in the fruit fly, *Drosophila*, than in any other organism. In *Drosophila*, homeotic transformations can be caused by mutations in any of a few dozen genes; two clusters of homeotic genes, the bithorax complex (BX-C [3]) and the Antennapedia complex (ANT-C [4]) have been the focus of most of the research to date. Each of the genes in the two complexes is expressed in a certain region of the *Drosophila* embryo during early development, and correspondingly that part of the embryo follows a different developmental pathway in the absence of that particular gene function (reviewed in Refs. 5–9). Because the genes appeared to have similar functions but in different places, Lewis (Ref. 3 and references therein) proposed that the genes might all have evolved from an

Abbreviations: *Ubx, Ultrabithorax*; *Antp, Antennapedia*; bp, base pair; *fiz, fushi tarazu*; *prd, paired*; eve, *even-skipped*; *AbdB, Abdominal B*; *gsb, gooseberry*; *abd A, abdominal A*; *Scr, Sex combs reduced*; *Dfd, Deformed*; *lab, labial*; *pb, proboscipedia*; *bcd, bicoid*; *zen, zerknüllt*; *hb, hunchback*.

Correspondence: M.P. Scott, Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO, 80309-0347, U.S.A.

ancestral gene by duplication and divergence. Diversification of the homeotic genes could have led to the evolution of insects, with their varied body segments, from annelid-like ancestors that had simpler repeating patterns of body segments. Lewis was therefore predicting that the genes of the BX-C and ANT-C would have

TABLE I

A compilation of homeodomain amino acid sequences

The homeodomain sequences are grouped into families of related sequences. The members of each class are shown compared to the *Antp* homeodomain. Dashes indicate identity with the corresponding residue in the *Antp* homeodomain. An X indicates that the residue has not yet been published, or in the case of MATa1 that the protein ends. The residues that are absolutely conserved within each class are shown below each set as the consensus sequence. It is possible that some of the classes will eventually be further subdivided. For example the cp11, xlhbox5 and zf25 homeodomains might constitute a class.

ANTP	*Organism*	*Reference*	*Synonyms*
Antp : RKKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEN	fly	15,17	--
abdA : P.....P.....L	fly	a	lab-2
c1 : X.....Y.....	human	111	--
c8 : X-R...I-S.....N.....S	human	111	--
cp11 : XG--S-TS.....NN...N.....DS	human	116	--
fts : DS--T.....I.....D--N--S-S.....S--DR	fly	120	--
hbl : C...C.....LS-L-C.....Y...S	sea urchin	199	--
hox11 : D.....TL.....H	mouse	91,200	m8
hox12 : HGR.....T.....N.....	mouse	91,200	m5
hox22 : SCR.....Y.....S	mouse	201,101	--
hox23 : D.....Y.....T.....	mouse	202,101	--
hox81 : D-R...I-S.....N.....S	mouse	106	--
hu2 : PTAG.....Y.....S	human	111	--
huhox12 : XGR.....N.....	human	116	--
huhox22 : XTA.....Y.....S	human	116	--
rlb : .....Y.....GV.....	rat	203	--
r5 : DG.....AV.....H	rat	204	--
Scr : -T--Q-TS.....H	fly	77,205	--
Ubx : L-R.....T-H.....M.....L...I	fly	17	--
Xhox36 : D.....H	frog	208,208	xlhbox3
xlhbox1 : D-R...I-S.....N.....S	frog	29,206	Xeb-1,AC-1
xlhbox2 : D.....T.....	frog	25,206,207	mm-3
xlhbox5 : DG--S-TS.....NN...N.....DT	frog	206	--
zf25 : XG--S-TS.....NN...N.....DS	zebrafish	209	--
Consensus: .....R--Y-R-QTLELEKEFH-N-Y-TRRR.....ERQIKIWFQNRRMK-KK--			
10 20 30 40 50 60			

DFD	*Organism*	*Reference*	*Synonyms*
Antp : RKKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEN			
Hu1 : DG--A-TA.....S.....D-	human	110,111	c10
c13 : XP--S-TA---Q-V.....T---S.....DH	human	111	--
cp10 : XP--S-AA---Q-V.....Y.....S---S.....DH	human	116	--
Dfd : -P--Q-TA---H-I.....Y.....T-V-S.....D-	fly	77,22	--
hox1.3 : -G--A-TA.....S.....D-	mouse	97,102	m2
hox1.4 : -P--S-TA---Q-V.....T---S---V.....DH	mouse	94,95	HBT-1,MH-3,Hox1.3
hox2.1 : DG--A-TA.....S.....DN	mouse	100,21	Mu-1,H24.1
hox2.6 : -P--S-TA---Q-V.....Y.....V.....S.....DH	mouse	103	--
hox5.1 : -P--S-TA---Q-V.....T---P.....DH	mouse	210	--
huhox1.3 : XG--A-TA.....S.....D-	human	211	--
huhox1.4 : XP--S-TA---Q-V.....T---S---V.....DH	human	116	--
huhox2.6 : XP--S-TA---Q-V.....Y.....V.....S.....DH	human	116	--
r3 : PGPARGVANG-Q-V.....Y.....S---S.....DH	rat	203	--
Xhox1a : -A--S-TA---Q-V.....Y.....V---T-R-S.....DH	frog	87	--
Xhox1b : DG--A-TA.....T---S.....D-	frog	87,207	xlhbox4
Consensus: .....A--R-Q-LELEKEFH-NRYLTRRRR-EIAH-L-L-ERQ-KIWFQNRRMKWKKD-			
10 20 30 40 50 60			

labial	*Organism*	*Reference*	*Synonyms*
Antp : RKKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEN			
hox1.6 : QPNAV-TNF-TK-LT.....K....A--V---AS-Q-N-T-V.....Q--RE	mouse	96	--
labial : TNNS--TNF-NK-LT.....A.....NT-Q-N-T-V.....Q--RV	fly	72,73	--
Consensus: --N--RTNFT-KQLTELEKEFHFN-YLTRARR-EIA--LQLNETQVKIWFQNRRMKQKKR-			
10 20 30 40 50 60			

TABLE I (continued)

Abd B	*Organism*	*Reference*	*Synonyms*
Antp : ERKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEN			
AbdB : XVRKK-KP-SKF---------L--A-VSKQK-W-L-RN-Q-----V----------N--NS	fly	89	iab-7
hox1.7 : STRKK-CP--KH---------L--M----D--Y-V-RL-N-----V----------M--I-	mouse	214	--
hox3.2 : STRKK-CP--K----------L--M----D--Y-V-RV-N-----V----------M--M-	mouse	212	--
huhox2.5 : XSRKK-CP--K----------L--M----D--H-V-RL-N-S---V----------M--M-	human	116	--
Consensus: --RKKR-PY-K-QTLELEKFFLFN-Y-----R-E-AR-L-L-ERQVKIWFQNRRMK-KK--			
10 20 30 40 50 60			

en	*Organism*	*Reference*	*Synonyms*
Antp : ERKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEN			
E30 : -E--P-TAFSAE-LAR-KR--AE-----E---QQLSRD-G---AE--------K-A-I--AS	honeybee	19	--
E60 : -E--P-TAFSGE-LAR-KR--AE-----E---QQLSRD-G-N-A--------K-A-I--AS	honeybee	19	--
en : DE--P-TAFSSE-LAR-KR--NE-----E---QQLSSE-G-N-A--------K-A-I--ST	fly	30,68	--
en1 : -D--P-TAF-AE-LQR-KA--QA---I-EQ--QTL-QE-S-N-S--------K-A-I--AT	mouse	33	Mo-en.1
en2 : -D--P-TAF-AE-LQR-KA--QT-----EQ--QSL-QE-S-N-S--------K-A-I--AT	mouse	33	Mo-en.2
inv : -D--P-TAFSGT-LAR-KH--NE-----EK--QQLSGE-G-N-A--------K-A-L--SS	fly	31	--
suhb-en : DE--P-TAFSAS-LQR-KQ--QQSN---EQ--RSL-KE-T-S-S--------K-A-I--AS	sea urchin	34	--
Consensus: --KRPRTAF---QL-RLK-EF----Y-TE-RR--L---L-L-E--IKIWFQNKRAK-KK--			
10 20 30 40 50 60			

eve	*Organism*	*Reference*	*Synonyms*
Antp : ERKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEN			
eve : SVR-Y-TAF--D-LGR-----YKEN-VS-P--C-L-AQ-N-P-ST--V--------D-RQR	fly	66,67	572
xhox-3 : XMR-Y-TAF--E-IAR-----YREN-VS-P--C-L-A--N-P-TT--V--------D-RQR	frog	d	--
Consensus: --RRYRTAFTR-Q--RLEKEFY-ENYVSRPRRCELAA-LNLPE-TIKVWFQNRRMKDKRQR			
10 20 30 40 50 60			

prd	*Organism*	*Reference*	*Synonyms*
Antp : ERKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEN			
gsbBSH4 : KQR-S-T-F-AE-LEA--RA-SRTQ-PDVYT-E-L-QTTA---AR-QV--S---ARLR-HS	fly	80	--
gsbBSH9 : KQR-S-T-FSND-IDA--RI-ARTQ-PDVYT-E-L-QSTG--ARVQV--S---ARLR-QL	fly	80	--
prd : KQR-C-T-FSAS-LD---RA-ERTQ-PDIYT-E-L-QRTN---AR-QV--S---ARLR-QH	fly	78	--
Consensus: KQRR-RTTF---Q---LER-F-RTQYPD-YTREELAQ-T-LTEAR-QVWFSNRRARLRK--			
10 20 30 40 50 60			

hox1.5	*Organism*	*Reference*	*Synonyms*
Antp : ERKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEN			
c13+1 : XS--V-TA--SA-LV----------C-P--V-M-NL-N---------------Y--DQ	human	116	--
hox4 : AS--A-TA--SA-LV----------FV-P--VQM-NL-N-S---------------Y--DQ	mouse	105	--
hox1.5 : SS----TA---P-LV----------M-P--V-M-NL-N---------------Y--DQ	mouse	89	Mo-10,HOX1.4
huhox2.7 : XS--A-TA--SA-LV----------FV-P--V-M-NL-N-S-------------Y--DQ	human	116	--
r6 : RHP--CTA--SA-LV----------C----V-M-NL-N---------------Y--DQ	rat	203	--
Consensus: ---R--TAYT--QLVELEKEFHFNRY--R-RRV-MANLLNL-ERQIKIWFQNRRMKYKKDQ			
10 20 30 40 50 60			

hox2.4	*Organism*	*Reference*	*Synonyms*
Antp : ERKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEN			
hox2.4 : G-R------S----------L--P----K----VS---G-----V----------	mouse	101	--
hox3.1 : G-RS-----S----------L--P----K----VS---G-----V----------	mouse	212,213	EA,m31
huhox2.4 : X-R------S----------L--P----K----VS---G-----V----------	human	116	--
huhox3.1 : X-RT-----S----------L--P----K----VS---G-----V----------	human	116	--
r1a : RTQ------S----------L--P----K----VS---G-----V----------	rat	203	--
r4 : G-RS-----S----------L--P----K----VS---G-----V----------	rat	203	--
Consensus: ----GRQTYSRYQTLELEKEFLFNPYLTRKRRIEVSHALGLTERQVKIWFQNRRMKWKKEN			
10 20 30 40 50 60			

TABLE I (continued)

	POU	Organism	Reference	Synonyms
Antp	ERKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEN			
unc86	DK--K-TSIAAPEKR---QF-KQQPRPSGE-IAS--DR-D-KKNVVRV--C-Q-Q-Q-RDF	nematode	43	--
Pit-1	RKRKR-T-ISIAAKDA--RH-GEHSKPSSQEIMRMAEE-N-EKEVVRV--C---QRE-RVK	rat	62,153	--
OCT-1	R-RKK-TSIETNIRVA---S-LE-QKP-SEEITM--DQ-NMEKEV-RV--C---Q-E-RI-	human	60	OTF-1,OBP100
OCT-2	R-RKK-TSIETNVRFA---S-LA-QKP-SEEILL--EQ-HMEKEVVRV--C---Q-E-RI-	human	57,60a	OTF-2
Consensus:	-----RT-I-------LE--F-----P----I------L---K-V-RVWFCN-RQ--KR--			
	↑ 10 ↑ 20 ↑ 30 ↑ 40 ↑ 50 ↑ 60			

	Unclassified homeodomains	Organism	Reference	Synonyms
Antp	ERKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEN			
bcd	RPR-T-T-F-SS-IA---QH-LQG----AP-LADLSAK-A-GTA-V----K---RRH-IQS	fly	78	--
cad	TKDKY-VV--DF-R------YCTS--I-I--KS-L-QT-S-S---V--------A-ERTS-	fly	63,65	S67
cut	PS-KQ-VLFSEE-KEA-RLA-ALDP-PNVGTIEFL-NE-G-AT-T-TN--H-H--RL-QQV	fly	36	--
HS.0	K-SWS-AVFSNL-RKG--IQ-QQQK-I-KPD-RKL-AR-N--DA-V-V---------RHTR	fly	37	--
ks	-AR-L-TA--NT-L----------K--C-P--V---AL-D-----V-V---------H-RQT	human	215	--
mab5	KS--T----S-S----------YHK----K--Q--SET-H-----V----------H---A	nematode	216	--
JML1001	KMR-A-TAF-YE-LVRV-NK-LTS---SVVE-LNL-IQ-Q-S-T-V--------T----H-	nematode	b	--
mec3	K-RQP-T-IKQN-LDV-NEM-SNTPKPSKHA-AKL-LETG-SM-V-QV------S-ERRLK	nematode	29	--
ro	KQR-Q-T-FSTE---R--V---R-E-IS-S--F-L-HT-R---T----------A-D-RIE	fly	40,41	--
sem1	RV-LK-TAF-SV-LV---N--KS-M--Y-T------QR-S-C---V----------F--DI	fly	71	S60
sem2	KS--S-TAFSSL-LI---R---L-K--A-T-----SQR-A-----V----------L--ST	fly	71	--

	yeast	Organism	Reference	Synonyms
Antp	ERKRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALCLTERQIKIWFQNRRMKWKKEN			
MATa1	KSPK-KSSISPQARAF--QVKRRKQS-NSKEKE-V-KKCGI-PL-VRV--I-K--RS-XXX	yeast	191,217	--
MATa2	Y-GHRFTKENVRILESWFAKKNIEP--DTKGLENLMKNTS-SRI---N-VS---R-E-TIT	yeast	191	-- (see note c)
PHO2	KQRPK-TRAKGEALDV-KRK-EI-PTPSLVE-KK-SDLIGMP-KNVR-------A-LR-KQ	yeast	218	--
mat2-P	TVRQQCSKC-KPHLMRWLLLHYD-P-PSNSEFYDLSA-TG--RT-LRN--S---R	yeast	46	

[a] W. Bender, personal communication
[b] Hawkins and McGhee, personal communication
[c] The homology of MATα2 to *Antp* is increased if the triplet NIE is deleted from the MATα2 sequence.
[d] Ruiz i Altaba and Melton, submitted

related structures. The two gene complexes have been cloned [10–14] and, indeed, Lewis was correct, although not in quite the way anyone could have guessed. A gene of the BX-C (*Ultrabithorax* (*Ubx*)) and a gene of the ANT-C (*Antennapedia* (*Antp*)) were found to contain a similar sequence of about 180 base pairs (bp) that has been named the homeobox [15–19]. At the DNA sequence level the two sequences are 74% the same. The homeobox constitutes part of the protein coding sequence of each gene and the corresponding 60 amino acid part of each protein is called the homeodomain. The *Antp* and *Ubx* homeodomains are identical at 54 amino acids out of 61 (88%) (Table I). The observation of greater conservation of the protein sequence than of the DNA sequence suggested that it is the protein sequence that is being selected and maintained during evolution. Homeodomains form parts of much larger proteins, and are usually located near the C-terminus of the proteins. Generally, there is little similarity between the primary sequences of the proteins outside of the homeodomain. Some exceptions to this have been noted, however (e.g., Refs. 20–23; W. Bender, personal communication), such as the sequence MXSYP at or near the N-terminus and the sequence YPWM positioned upstream of the homeodomain.

The homeotic genes direct the development of the different segmental structures in the epidermis, mesoderm and the nervous system. An earlier acting class of genes directs the formation of the segmental divisions of the embryo. One of these 'segmentation genes', the *fushi tarazu (ftz)* locus, is in the ANT-C [24]. An embryo homozygous for a *ftz* mutation has half the normal number of body segments. Surprisingly the *ftz* gene was found to contain a homeobox [15,17] which immediately suggested that *ftz* and the homeotic genes, even though their phenotypes are very different, encode related proteins and may use similar molecular mechanisms.

Drosophila has been popular as an experimental organism for developmental geneticists due to an 80 year investment in its genetics and to its rapid and accessible development. Because so little is understood about the genetic control of development in any organism, it was reasonable to study insect development, with its experimental advantages, and to hope that the results would provide some guidance for studies

in classes of animals less amenable to genetic analysis, including mammals. The homeobox is one example of the fulfillment of this hope. Homeoboxes quite closely related to the first known fly homeoboxes were found by DNA cross-hybridization in mice, humans, chicken, *Xenopus*, and earthworms [16]. Subsequently, homeoboxes have been found in many higher eukaryotes including representative arthropods, annelids, ascidians, echinoderms, brachiopods, tapeworms, molluscs and chordates (e.g. Refs. 25–27), but not in coelenterates, nematodes, sponges flatworms, slime molds, fungi, bacteria, platyhelminths or aschelminths. The recent detection of homeoboxes in nematodes (Refs. 28, 43, 216; Hawkins, N. and McGhee, J., personal communication) suggests that some of the failures to detect homeoboxes by DNA cross-hybridization in certain species may be due to technical limitations rather than to the actual absence of homeoboxes from some of the phyla, but this remains to be shown. The first reported sequence of a homeobox from a higher eukaryote, *Xenopus*, revealed striking sequence conservation, the strongest conservation being at the protein sequence level [29]. The matches with the *Antp*, *Ubx* and *ftz* homeodomain sequences were 55/60, 51/60 and 49/60, respectively.

Homeoboxes have now been found in over 20 *Drosophila* genes. Most of these genes are known to regulate development. Some were cloned using homeobox DNA cross-hybridization; in other cases the homeobox was discovered only after the gene had been cloned in other ways. In addition, more than 50 homeobox sequences have been obtained from non-*Drosophila* species. In most cases the functions of these genes are unknown, but we describe below several cases in which the regulatory functions of non-*Drosophila* homeobox genes are known or can be inferred. As we will describe, all homeodomain-containing proteins that have been localized have been found in the nucleus, and there is similarity between the homeodomain and known DNA binding proteins and transcription factors. The current hypothesis is that a homeodomain is indicative of a role in transcriptional regulation, but such a role has only been firmly demonstrated in a few cases.

Currently important questions about homeoboxes are whether the homeodomain is a DNA binding domain in vivo, whether the homeodomain-containing proteins are transcription factors, what the functions of homeobox genes are in species other than *Drosophila* and how the *Drosophila* gene network that controls development makes use of its many homeoboxes. In this review we will focus on two areas: the common structural characteristics of homeoboxes and homeodomains as deduced from sequence comparisons, and the molecular and developmental functions of homeodomain-containing proteins.

II. Classes of homeoboxes

II-A. A sequence compilation

Homeobox sequences are being reported at a very rapid rate, and this has led to difficulties with nomenclature and to some questions as to what constitutes a homeobox. The first homeoboxes found were closely related over a 180 bp region, but some sequence homology has been found in flanking regions of the DNA (and protein) as well. Lower stringency hybridizations, as well as the identification of homeoboxes in sequences of genes that had been isolated for other reasons, have led to the discovery of much more highly divergent homeoboxes (and homeodomains). A table of 87 homeodomain sequences (Table I) demonstrates that many of the proteins can be grouped into related classes. The sequences are aligned at an N-terminal position of (most commonly) a glutamate or aspartate, and 61 amino acids are shown for each protein. The most distantly related of the higher eukaryote sequences in this table are identical or conservatively substituted at 30 of the 61 amino acids.

21 homeodomains have been identified in *Drosophila*. 20 human and 19 mouse homeodomains have already been found. The other 27 homeodomains were isolated from *Xenopus* (seven), rats (seven), nematodes (four), yeast (four), sea urchin (two), honey bees (two) and zebrafish (one). Although all 87 homeodomains shown in Table I are clearly related to each other, they can be put into sets that have more closely related sequences. From the 83 sequences (excluding the yeast homeodomains) in Table I, we have sorted the homeodomains into ten classes, with 14 sequences not in any group (Table I). A consensus sequence is shown for each group of homeodomains (Table IIA). Several of the groups contain sequences from multiple species, suggesting that the common features among the homeodomains in a group have been evolutionarily conserved. The degrees of similarity between the different fly homeodomains are shown in Table III. The first class to be identified has been referred to as the *Antennapedia* class homeodomains. There are currently 24 members of this class, and all are identical to the *Antennapedia* homeodomain at more than 50 out of 61 positions.

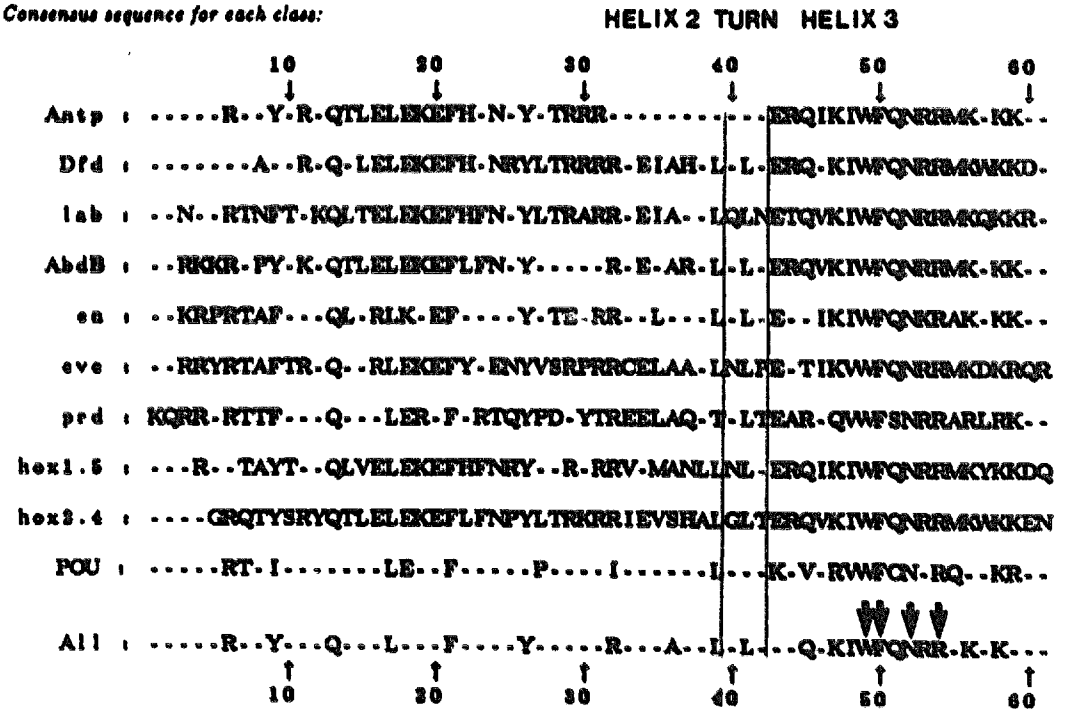
One class of homeodomains that stands out is the *engrailed* group, of which there are two in *Drosophila* (*engrailed* and *invected*), two in honeybee, two in mouse, and one in sea urchin (see references in Table I). The *engrailed* class sequences are related to each other by having sequence identities at 70–80% of the amino acids and by the position of an intron relative to the homeobox [30,31]. The *engrailed* class sequences are all only about 45% conserved at the amino acid level with the *Antennapedia* class sequences. In addition, at least some

TABLE II

Homeodomain consensus sequences and variability

(A) The consensus sequence for each class of homeodomains, from Table I, is shown here, as is the consensus for all of the 83 higher eukaryotic sequences ('All'). Only absolutely conserved residues are shown for the consensus sequences for the individual classes, while the overall consensus includes highly conserved residues that are not absolutely conserved. Yeast sequences were not included in the compilation for 'All'. A dash indicates that the position is variable. The arrows indicate the amino acids that are absolutely conserved in all of the higher eukaryotic homeodomain sequences; three of the four amino acids in this category are also conserved in the yeast sequences. (B) The consensus from all of the homeodomains (from (A)) is at the top, with the three predicted α helices and the turn indicated. The *Antp* sequence is shown for reference purposes. Helix 3 is the putative recognition helix. Below the consensus are shown all of the amino acids that appear at each of the 61 positions in the 83 higher eukaryotic sequences. The least variable region corresponds to the recognition helix.

A. Consensus sequence for each class:



B. Amino Acids at Each Position

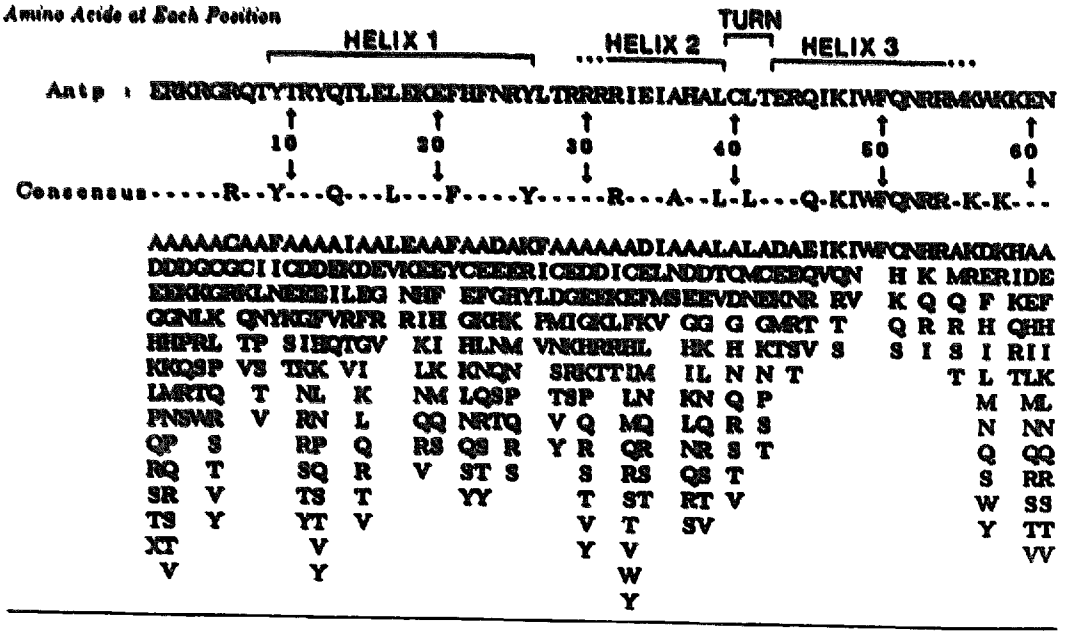


TABLE III

Sequence homologies among the Drosophila homeodomains

The number of amino acids that are identical, out of 61, is shown for each pair of sequences. The second number in each box is the number of amino acids in common if conservative substitutions are allowed (i.e., S for T, E for D, R for K, and I for V or L).

	abdA	AbdB	Antp	bcd	cad	cut	Dfd	en	eve	ftz	BSH4	BSH9	H20	inv	lab	prd	ro	Scr	Ubx	zen1	zen2
abdA	61-61	36-46	58-58	25-30	32-40	18-23	47-50	30-34	30-36	49-53	22-28	21-29	24-33	29-33	40-42	22-29	34-40	54-55	53-55	36-40	37-42
AbdB	36-46	61-61	35-45	23-30	30-37	18-21	31-43	24-33	29-37	33-45	21-28	22-27	25-30	25-32	32-38	22-28	33-39	35-45	35-43	32-39	31-40
Antp	58-58	35-45	61-61	25-30	32-40	17-21	49-52	30-35	30-36	50-55	22-28	20-28	24-33	30-34	40-41	22-29	34-40	56-57	54-55	36-39	37-42
bcd	25-30	23-30	25-30	61-61	22-27	17-22	26-32	27-34	25-32	26-31	23-30	23-29	23-32	27-34	26-30	23-31	24-31	25-31	25-28	26-32	27-38
cad	32-40	30-37	32-40	22-27	61-61	15-23	33-40	24-32	24-33	33-39	20-29	19-29	24-32	23-31	30-34	20-29	27-35	31-38	31-38	28-34	26-36
cut	18-23	18-21	17-21	17-22	15-23	61-61	18-22	20-24	19-25	19-23	24-28	24-31	16-22	20-23	17-21	22-27	19-23	18-23	17-22	17-22	20-24
Dfd	47-50	31-43	49-52	26-32	33-40	18-22	61-61	32-38	31-39	45-49	21-30	19-30	21-32	32-37	39-42	20-30	33-41	50-53	43-47	38-42	37-45
en	30-34	24-33	30-35	27-34	24-32	20-24	32-38	61-61	27-36	30-35	23-31	22-31	21-29	52-57	30-35	23-29	31-37	31-36	28-33	30-34	36-41
eve	30-36	29-37	30-36	25-32	24-33	19-25	31-39	27-36	61-61	30-37	25-31	24-32	26-33	27-34	30-36	26-32	34-40	31-37	31-35	31-37	29-37
ftz	49-53	33-45	50-55	26-31	33-39	19-23	45-49	30-35	30-37	61-61	20-29	18-29	23-33	28-34	38-41	20-29	33-40	48-55	46-51	36-40	35-44
BSH4	22-28	21-28	22-28	23-30	20-29	24-28	21-30	23-31	25-31	20-29	61-61	49-55	22-26	23-29	22-28	50-53	29-34	22-29	24-28	22-29	27-35
BSH9	21-29	22-27	20-28	23-29	19-29	24-31	19-30	22-31	24-32	18-29	49-55	61-61	25-27	22-29	21-30	49-52	27-34	20-29	22-29	21-30	27-33
H20	24-33	25-30	24-33	23-32	24-32	16-22	21-32	21-29	26-33	23-33	22-26	25-27	61-61	21-29	24-32	24-27	24-31	23-32	23-30	23-32	28-35
inv	29-33	25-32	30-34	27-34	23-31	20-23	32-37	52-57	27-34	28-34	23-29	22-29	21-29	61-61	30-34	23-29	30-35	31-35	28-32	29-33	34-39
lab	40-42	32-38	40-41	26-30	30-34	17-21	39-42	30-35	30-38	38-41	22-28	21-30	24-32	30-34	61-61	21-28	32-38	40-41	37-39	38-39	37-40
prd	22-29	22-28	22-29	23-31	20-29	22-27	20-30	23-29	26-32	20-29	50-53	49-52	24-27	23-29	21-28	61-61	28-32	23-31	24-29	23-31	28-34
ro	34-40	33-39	34-40	24-31	27-35	19-23	33-41	31-37	34-40	33-40	29-34	27-34	24-31	30-35	32-38	28-32	61-61	35-42	35-39	29-38	32-40
Scr	54-55	35-45	56-57	25-31	31-38	18-23	50-53	31-36	31-37	48-55	22-29	20-29	23-32	31-35	40-41	23-31	35-42	61-61	50-52	37-40	38-44
Ubx	53-55	35-43	54-55	25-28	31-38	17-22	43-47	28-33	31-35	46-51	24-28	22-29	23-30	28-32	37-39	24-29	35-39	50-52	61-61	35-40	36-41
zen1	36-40	32-39	36-39	26-32	28-34	17-22	38-42	30-34	31-37	36-40	22-29	21-30	23-32	29-33	38-39	23-31	29-38	37-40	35-40	61-61	43-47
zen2	37-42	31-40	37-42	27-38	26-36	20-24	37-45	36-41	29-37	35-44	27-35	27-33	28-35	34-39	37-40	28-34	32-40	38-44	36-41	43-47	61-61

of the *engrailed* sequences have extended homology adjacent to the 61 amino acids: 26 amino acids upstream and 20 amino acids downstream [32–34].

Other homeoboxes fall outside the *Antennapedia* and *engrailed* classes. The *paired* (*prd*), *labial* (*lab*), *even-skipped* (*eve*) and *Abdominal B* (*AbdB*) classes have three, two, two and four members, respectively. The *prd* and two *gooseberry* (*gsb*) region genes (BSH4 and BSH9) have an 18 amino acid extension of the homeodomain region of homology at the N-terminus of the 61 amino acid sequence shown in Table I [35]. The POU class has one member from *Caenorhabditis elegans* and three mammalian members. The grouping of these four genes into a class is substantiated by the finding that all four proteins have two additional regions of conserved sequence upstream of the homeodomain (now called the POU domain); the two additional regions are even more conserved than the homeodomain [36]. The two remaining classes (hox2.4 and 1.5) contain only mammalian homeodomains at the present time. Many of the divergent homeobox sequences cannot readily be grouped into any related subclasses. Some of them such as *H2.0*, cannot readily be detected by cross-hybridization to a homeobox probe (only the *Scr* probe, but not *Ubx* or *Antp*, can be used to detect *H2.0*, Ref. 37), so other still more divergent homeoboxes probably remain to be found. The structural relatedness of homeodomain subclasses may not reflect any functional similarities between members of a given class, but rather the evolutionary history of the sequences.

In spite of the differences between classes of homeodomains, all of the homeodomains in Table I (excluding yeast homeodomains) exhibit striking similarities in primary sequence, even without the introduction of gaps in any of the domains. These structural similarities in large part define the homeodomain, and therefore are worthy of detailed discussion. Four amino acid residues are conserved in all non-yeast homeodomains, all in the C-terminal third of the domain (see Tables I and II), and three of the four are also conserved in the yeast sequences. The invariant residues are tryptophan, phenylalanine, asparagine and arginine at positions 49, 50, 52 and 54, respectively. Another eight positions are very highly conserved, though not invariant. These are positions 6(arginine/glycine), 13(usually glutamine), 17(leucine/valine), 21(phenylalanine/tyrosine), 41(leucine/asparagine), 46(isoleucine/valine), 56(lysine/arginine), and 58(lysine/arginine). The extreme conservation of these 13 residues, together with the high degree of conservation at other positions (see Tables I and II), is characteristic of most of the homeodomains identified to date. Certain other features are characteristic of homeodomains and serve as an aid in determining whether a homeodomain-related sequence is in fact a true homeodomain. One of the distinguishing features is the considerable conservation of predicted secondary structure in homeodomains, even in regions that tend to be relatively variable in primary sequence. The predominantly α-helical nature of specific regions of the homeodomain has important functional implications, and is therefore discussed in detail in a later section. A second distinguishing characteristic is the nuclear location of all homeodomain-containing proteins examined to date. This subcellular distribution is consistent with their proposed role as transcriptional regulators (see later discussion). Thus, the subcellular location and secondary structure, as well as the primary sequence of a given protein are important factors in the identification of new homeodomains.

What sorts of genes contain homeoboxes? The *Drosophila* genes (see Table I for references) include homeotic genes that control segmental differentiation (*lab, Dfd, Scr, Antp, Ubx, abdA, AbdB*), segmentation genes that control the division of the embryo into segments and that in some cases control the homeotic genes (*ftz, eve, prd, en, gsb, inv*), genes involved in dorsal/ventral differentiation (*zen1* and *zen2*), genes that act maternally to control anterior–posterior polarity of the embryo (*cad, bcd*), a gene (*cut*) that controls cell determination in the peripheral nervous system and also functions in leg, head and wing development [38], and a gene (*ro*) that functions in eye development (39–41]. One *Drosophila* gene of unknown function, *H2.0*, is expressed in a tissue-specific rather than position-specific pattern in the developing fly [37]. Four nematode homeobox genes have been identified. The function of one is unknown, another, (*mec3*) controls differentiation of touch receptor neurons [28], and one (*mab5*) controls cell fates in the posterior of the worm functions [42]. The fourth (*unc86*) is a cell lineage gene [44]. The yeast genes include three that control mating type (*MAT*α*2* and *MAT***a1**) (reviewed in Ref. 45), and the fission yeast gene *mat2-P* [46], and also a gene (PHO2, also called BAS2) that is a transcriptional activator of an acid phosphatase gene [47,48]. The expression patterns of mouse homeobox genes suggest roles in controlling development (reviewed in Refs. 49–55). Although the functions of most of the vertebrate homeobox genes remain unknown, Pit-1 and the OCT factors have been shown to be transcriptional regulators [56–62].

To what extent does the presence of a homeobox correlate with interesting regulatory genes? One way to answer the question is to ask how often a gene that has been cloned using a homeobox DNA probe turns out to be a known gene of interest. The yeast genes and two of the three nematode genes were first identified through their interesting phenotypes and then found to have homeobox-like sequences. This is also true of many of the *Drosophila* genes. In contrast, the *cad* gene was first cloned through its homeobox homology, then found to have an interesting distribution [63,64], and then found

to have an interesting phenotype [66]. Previously known loci of importance that have been cloned through cross-hybridization with a homeobox DNA probe include *eve* [66,67], and *en* [68]. In addition, presumed protein-coding regions for genes of interest were located using homeobox probes for *abdA* and *AbdB* [69], *zen1* and *zen2* [70,71], and the putative *lab* homeobox (Refs. 72 and 73; T. Kaufman, personal communications). One *Drosophila* homeobox (*H2.0*) has been found for which there is no known function and no genetic linkage to a gene of known function, but the distribution of the transcripts suggests a role for the gene in tissue differentiation [37]. Therefore there has been excellent success in finding genes that appear to regulate development simply by using a homeobox probe to screen genomic libraries.

Highly diverged homeoboxes cannot be so easily isolated. Divergent homeodomains in yeast proteins (MATa1 and MATα2) were found by protein structure similarity soon after the discovery of homeoboxes in *Drosophila* genes involved in the control of development. While these homologies provided one of the first clues as to the function of the homeodomain, they also serve to illustrate a current topic of debate in the field. Should highly divergent sequences still be considered to be homeodomains? In light of newly reported, highly divergent homeodomains, can the presence of a homeodomain in a protein still be considered indicative of a developmental function, or has the definition of a homeodomain become too vague? The homeodomains found in the yeast DNA-binding proteins MATa1 and MATα2, although highly diverged from the majority of homeodomains identified to date, are clearly structurally related to the homeodomains found in higher eukaryotes. Since MATa1 and MATα2 play roles in the control of yeast mating type, they support the view that the presence of a homeobox is an indicator of developmentally important genes. Homology between another two yeast proteins (PHO2 and ARD1) and homeodomains has recently been reported. ARD1 is a gene that controls the switch between the mitotic cell cycle and alternative cell fates such as stationary phase of sporulation [74]. The sequence of ARD1, while somewhat related to certain homeodomains, clearly does not fit the pattern of conservation observed in the homeodomains listed in Table I. Therefore, while this protein may be distantly related to homeodomain-containing proteins, it does not contain a true homeodomain by the criteria of the conserved amino acids described above. In contrast, PHO2 (a transcriptional activator of an acid phosphatase gene) exhibits significant homology to the *prd* and *en* homeodomains (37% in both cases). Without the introduction of gaps in its primary sequence, all four of the invariant positions found in all of the non-yeast homeodomains in Table I are conserved in the PHO2 homeodomain. Six of the eight highly conserved positions are also conserved in PHO2. Thus, PHO2 clearly contains a homeodomain (as defined by sequence), although it plays no obvious developmental role. It is worth noting, however, that the role of this gene in known processes does not preclude the possibility that it also functions in yeast processes that may be closer to developmental functions than is the cell metabolic regulatory function. Other proteins may be found to have similarities to only part of the homeodomain. An example of a protein that has a possible helix-turn-helix sequence (see below) but is not otherwise similar to a homeodomain is *fs*(1)*K*10 [75], a gene that controls dorsal/ventral polarity in the *Drosophila* embryo. We agree with Prost et al. [75] that the *fs*(1)*K*10 protein, and others with similarly short sections of sequence that are reminiscent of homeodomains should not be classified as homeodomains.

The sequence similarity of ARD1 and PHO2 to homeodomains demonstrates the need for caution in defining what constitutes a homeodomain, and in interpreting the functional significance of this domain. The first homeodomains were identified by comparison of the primary sequence of functionally related proteins. The fact that so many developmentally important genes have been found to contain homeoboxes does suggest that the homeodomain is extensively used in genetic systems controlling development. However, many homeobox genes have been isolated by virtue of their DNA sequence homology to *Drosophila* homeotic and segmentation genes, and thus might be expected to be functionally related as well. Since we currently have no idea how many homeoboxes are present in the genome of any organism, including *Drosophila*, it is possible that a great variety of homeodomain-containing proteins exist, and may function in regulating the activity of a broad spectrum of target genes. Thus, in the absence of data concerning the function of a given homeodomain-containing protein, a developmental role should not be assumed unless the homeodomain homology is quite high or the homology extends to non-homeodomain regions of developmentally important proteins. At present, it does seem safe to conclude that a protein containing a homeodomain probably functions as a DNA-binding protein that regulates the expression of other genes, though not necessarily in a developmental context.

II-B. Genomic organization and the evolution of homeoboxes

One of the striking features of the homeotic genes of *Drosophila* is the clustering of many of them into the two complexes. The BX-C contains the three homeoboxes of the *Ubx*, *abdominal A* (*abdA*), and *AbdB* homeotic genes [15,17,69]. The ANT-C (reviewed in Ref. 76) contains the *Antp*, *Sex combs reduced* (*Scr*),

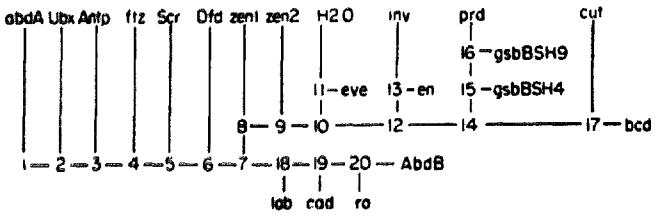


Fig. 1. Evolutionary tree of *Drosophila* homeodomains. The tree was generated by the program 'protpars' of the Phylip package (Phylip was generously provided by J. Felsenstein). The tree was generated as the most parsimonious way for the protein sequences to have evolved taking into account relationship between codon sequences. The numbers represent inferred intermediates in the evolution of the sequences; the sequences of the intermediates are shown in Table IV. This is not the only possible tree, but it is most parsimonious among the trees generated and it is the one most frequently generated by the analysis when the order in which sequences are added is varied. There is no significance to the length of the lines connecting the sequences, and parts of the tree could be rotated around any of the branch points.

Deformed, (*Dfd*), *labial* (*lab*), and *proboscipedia* (*pb*) homeotic genes, each with its homeobox (Refs. 15, 17, 69, 72, 76, 77; T. Kaufman, personal communication). In addition there are at least four other genes with homeoboxes in the complex: *ftz* [15,17], *bicoid* (*bcd* [78]), and *zerknüllt* (*zen*) genes 1 and 2 [71]. While the *ftz* homeobox (and homeodomain) sequence is closely related to the homeotic gene sequences, the other three are not (Table I). Therefore the simplest idea – that the ANT-C arose through duplication and divergence of an ancestral gene containing a homeobox – is not supported in an obvious way by the presence of the divergent homeoboxes in the complex.

Most of the other *Drosophila* homeoboxes are not clustered, two exceptions being the *engrailed* and *invected* genes [30,31,68] and the two *gooseberry* (*gsb*) region genes BSH4 and BSH9 [78]. In both of these cases the clustered genes and their products are closely related, are expressed in similar patterns, and seem very likely to have arisen from a common ancestor.

We have used the computer program 'protpars', from the 'Phylip' package, to generate a 'tree' of the *Drosophila* homeodomains (Fig. 1 and Table IV) that is based on relationships among the sequences. The tree is constructed with the assumption that the sequences evolved from a common ancestor, but it is of course possible that there has been some convergent evolution as well. The tree was constructed by computing a parsimonious way in which the sequences could have arisen by gradual changes, taking into account the number of changes required to convert one codon into another (reviewed in Ref. 79). Intermediate forms which link the actual sequences are generated by the program and are indicated as numbers 1–20 in the tree. The changes in sequence to generate the intermediates are shown in Table IV. The question marks in the sequences of the intermediates indicate cases where a change in sequence could have occurred at more than one place in the tree and there is not enough information to say when the change happened. For example, the proline (P) found in the first position in *abdA* could have arisen when the leucine (L) in *Ubx* became P in the transition to intermediate No. 2. Alternatively, there could have been an L in the first position of intermediate No. 2 and the change to P could have happened in the conversion of No. 2 into No. 1 or when No. 1 became *abdA*. Similar types of uncertainties apply to the other question marks in Table IV.

The intermediates shown in the tree are plausible, but variations on them could be closer to intermediates that actually arose during evolution, and there could have been multiple intermediates where only one is shown. The tree is 'unrooted': it contains information about the relations between sequences, without giving an explicit starting point. Intermediate No. 4 could have evolved into intermediate No. 5, or the opposite could have occurred. Only historical information could establish the actual events that occurred, and it is not clear that the necessary historical information can be obtained from doing present-day sequence comparisons.

The computer program generates somewhat different trees depending on the order in which the sequences are added. For this reason we used the program's capability to jumble the order in which it adds the sequences. 40 different orders of adding the sequences were used to increase the likelihood that the 'best' tree was found and to see what aspects of the tree were independent of how the sequences were added. The tree in Fig. 1, and slight variations on it, account for all of the best scoring trees (32 out of the 191 trees generated). The variations consist of: (i) the exchange of nodes 3 and 4 (a node is a branchpoint); (ii) the reordering of nodes 7, 8, 18 and 19 so that nodes 7 and 20 are connected and the new linkage pattern connects 7 to 19 to 18 to 8; and (iii) the placement of node 4 between nodes 6 and 7.

Many aspects of the tree are intuitively reasonable. The *gsb* and *prd* proteins, all involved in segmentation, share a related protein domain other than the homeodomain [80] and their clustering in the tree is therefore not surprising. *en* and *inv*, and *zen1* and *zen2*, are close together on their chromosomes, and each pair is expressed in similar patterns, so the homeodomain similarity is again expected. The homeotic genes in the two homeotic gene complexes are fairly closely positioned on the tree, except that *lab* and *AbdB*, which control formation of the extreme (and opposite) ends of the embryo, are placed on a divergent branch. The closeness of the homeotic gene complex sequences (see also Table III) may reflect the origin of the two complexes as a single complex. There appears to be one complex with characteristics of both of the *Drosophila* complexes in beetles [81].

TABLE IV

Changes in homeodomain sequences generated for the Drosophila evolutionary tree by the Phylip program

The first line, sequence No. 1, is an arbitrary starting point generated by the computer. Sequence No. 1 is shown at the top. Subsequent lines show the changes between the two sequences listed in the left-hand columns. Refer to Fig. 1 for the positions of the sequences in the evolutionary tree. Each line thereafter indicates the changes in the sequence, for example between No. 1 and abdA (second line) where two substitutions are made. The question marks, as explained in the text, indicate uncertainty as to when a change took place. The change of P (in *abdA*) into L (in *Ubx*) could have occurred between No. 1 and No. 2, or between No. 2 and *Ubx*, for example.

From	To	State at Node ("." means the same as in the nearest node with a lower number)
	1	?RKRGRQTYT R?QTLELEKE FHFNRYLTRR RRIEIAHALC LTERQIKIWF QNRRMKWKKE ?
1	abdA	P..... .F..... L
1	2	?..... .Y..... I
2	Ubx	L.R..... ....T.H..... ....M..... ....L... .
2	3	E..... .... ?
3	Antp	.... N
3	4	....R..... ....?..... .....? ?
4	ftz	DS..T..... ....I... ..D..N..S .S..... ......S..D R
4	5	....Q.T?.. ....?..... .....? ?
5	Scr	.T.....S.. .... .T..... .....E H
5	6	.......A.. ...I..... ....T.G .S..... .....D N
6	Dfd	.P..... .N.....D.. ..Y..... .....V .S..... ..... .
6	7	K...*...?. SF.L..... .....? .....?... .....V.... ......L..? ?
7	8	.?.....F. .V..V..... ..?.?....T .....QR.. ....? I
8	zen1	RV.LK..... S.....N. .KS.M..Y.. .....S .C..... ......F..D .
8	9	.?..S....S .....?. ..L.K..... ..... .T..... .....? T
9	zen2	.S..S..... SL..I...R. .....A.. .....S...A ....S .
9	10	.??..... ....EG..?. .?Q.....?? ..R.L.?... ...A.?.. . .....R .
10	11	.?R..... .....K. .?.?..?.?P .....A..N .....?.V.. ......?.R. R
11	H2.0	.RSWS.AV.. NL.RK...IQ .Q.Q..I.K. D..K..... ..D..V.... ......WRHT .
11	eve	SV..Y....T RD..GR.... .YKEN.VSR. ..C....Q.. .P.STI.... .....D..Q .
10	12	.??.?..... ??.....R. .Y..R...ER ..Q...??.. .....I.... ....?..... ?
12	13	EEK.P..... ??..AR.K.. .NE..... ...Q.S?E.. .N..... ..K.A....S ?
13	inv	.D..... GT.....H. .....K .....SG... ....S S
13	en	D..... SE..... ....SS... ....I..S T
12	14	.QR.?..T.. ??...A...A .S....??V. ?.E...??.. ....?..... ?...?R...Q S
14	15	....S..... A?..... ..RTQ.PD.Y T.....QRT. ....R.QV.. S...A..R.. .
15	16	..... .A..D..... .A..... L
16	prd	....C..... .S...E.... .E.....I. .....N ..... S..... H
16	BSH9	....S..... ND.I.....I ....S.. .....V.... S..... .
15	BSH4	....S....T .E..... .S..... .....T.A ..... S.....H S
14	17	RP..P..... ??.I....?. ...D..??.. ??.?..AK.. .AT.?..... ?...N..... .
17	cut	PSKKQ.VL.. EE.K...RL. .AL.P.PN.G TI.F..NE.. ...RT.TN.. N.N.....Q. V
17	bcd	....T....T SS..AE..QN .L.G..LTAP RLAD.S...A .G..QV.... K...R.N.I. S
7	18	TK?....??. N..... .....? .....?..R ....Q..? ?
18	lab	.NNSG..NF. .K..T..... .....A .....N..Q .N.T..... .....R V
18	19	..??...?Y. ...R..... ..?...I..R ..?.L.Q... ....?.E..S N
19	cad	..DKY.VV.. D..... YCTS....I. .KS.....S .S..... ....A..RTS .
19	20	??R??..T.S ?..T..... ..?.?..S.. ..?..... .T..... ....?.D..? ?
20	ro	-Q.RQ...F. TE...R..V. ..R.E....S ..F...E... ...T.I.... ....A...RI E
20	AbdB	-V.KK.KP.. K..... .LF.A.V.KQ K.W...RN.Q ....M.N..N S

The significance of the clustering of *Drosophila* homeotic genes in the ANT-C and BX-C is unknown, but Lewis observed [3] the curious fact that the order of homeotic genes along the chromosome in the BX-C corresponds to the order of the body segments that they affect. This observation has been extended by the work of Kaufman and his colleagues (e.g., Refs. 4, 76 and 82–84) to the ANT-C where the homeotic genes are also ordered along the chromosome according to their sites of function along the body of the fly. Thus the leftmost ANT-C gene, *lab*, acts in four segments of the head, the next gene, *pb*, acts in the posterior head segments (at

least in the development of the adult), the next gene, *Dfd*, is active just posterior to *lab*, and *Scr*, the next gene, is active in the posterior head and anterior thoracic segmental primordia. *Antp* itself, which is adjacent to *Scr*, acts in parts of all of the thoracic primordia (overlapping in its region of influence with *Scr*) and at a low level in the abdominal segments. Beyond *Antp* the next homeotic gene that is encountered, the *Ubx* gene, is a third of a chromosome arm away in the BX-C. *Ubx* functions in the posterior thorax (overlapping with *Antp*) and anterior abdomen. The *abdA* gene is adjacent to *Ubx* and functions in the central abdominal segment primordia. The last gene in the series, *AbdB*, functions in the most posterior abdominal regions. The exact details of where the genes are expressed and where they function will not be discussed here (reviewed in Refs. 5, 9 and 76).

It is curious that the order of the *abdA*, *Ubx*, *Antp*, *ftz*, *Scr*, *Dfd* and *zen* genes (and *bcd* if the tree is rotated at node 10) in the tree (Fig. 1) corresponds to their order along the chromosome. However, the two homeotic genes that function at the two opposite ends of the embryo, and which are located at opposite ends of the homeotic gene complexes, are placed by the analysis together on a branch of the tree that forks off between *Dfd* and the *zen* genes. Until more is known about the actual evolution of the genes, the importance of these observations cannot be readily assessed.

Three possible explanations of the ANT-C and BX-C gene organization are: (i) the correspondence is a coincidence. (ii) The arrangement of the genes is necessary for their differential expression along the embryo. Three chromosomal rearrangements that split apart the BX-C [85,86] have been used to show that the separated genes could still function (although there was some DNA in common between the separated pieces of the complex). (iii) The arrangement is an evolutionary remnant with no current relevance to gene function. In this last case the duplication and divergence of the genes from an ancestral gene could be viewed in a novel way: the genes could have evolved in the order that the body segments they control evolved. The evolution of insects from annelid-like ancestors could have occurred by head-controlling and tail-controlling genes forming first. The diversification of thoracic and abdominal segments would follow, controlled by later-evolving genes. What sort of homeobox gene would have existed prior to formation of the multiple homeotic genes? Perhaps the ancestral homeotic gene controlled differentiation events that are common to all segments.

Due to the lack of polytene chromosomes, much less is known about the genomic arrangement of non-*Drosophila* higher eukaryotic homeobox genes. However, it is striking that in several cases clustering has been observed that is reminiscent of the *Drosophila* gene complexes. In *Xenopus*, two homeoboxes have been found to be closely linked [87]. In both mice and humans there are clusters of homeobox genes. A system of nomenclature for the mouse genes has been proposed, and generally followed [88], in which the genes are called Hox genes. Each cluster of genes is given a number and each gene within the cluster the cluster number, a decimal point and its own number. So far the largest mouse cluster contains seven homeobox genes (Hox1.1–Hox1.7) in a 65 kb region on chromosome 6 [89–97]. In the mouse Hox2 locus on chromosome 11 there are four homeoboxes in a 20 kb region all transcribed in the same direction [98–101] plus a fifth for which the mapping has not yet been reported [10]. Sequence similarities among the Hox1 and Hox2 homeoboxes, together with the locations of mapped genes in the vicinity of the Hox1 and Hox2 loci, suggest that the two loci may have arisen by a large scale chromosome duplication event [97,102–104]. Single mouse homeoboxes have been found at the Hox4 locus on chromosome 12 [105], the Hox5 locus, and the Hox6 locus on chromosome 14 [106]. Two homeoboxes have been found at the Hox3 locus [107,108]. The two *engrailed* class mouse genes have been mapped to chromosomes 1 (En-1) and 5 (En-2) [109], so these two related genes are not linked on a chromosome.

The mouse Hox loci correspond to identified loci on human chromosomes [20,98,110–116]. 17 human homeoboxes have been cloned and sequenced. They are organized in four clusters with two homeoboxes within 20 kb on chromosome 2 [117], four within 23 kb on chromosome 7 [114], four in a 80 kb region of chromosome 12 [117], and seven in a 76 kb region on chromosome 17 all transcribed in the same direction [118]. For comparison, the *Drosophila* BX-C and ANT-C are each within about 300 kb regions. The nomenclature for human homeoboxes has unfortunately not been standardized as of the time this review was written.

It is not clear whether the clustering of homeobox genes in mammals is related to the clustering in *Drosophila*. The expression patterns of the mammalian genes are only partly known, but there is some evidence for a correlation between where the genes are expressed and their order along the chromosome [53]. One emerging generalization is that genes at the 5′ end of the Hox2 cluster (the genes are all transcribed in the same direction) are expressed in more posterior parts of the ectoderm and genes in more 3′ parts of the cluster have more anterior limits to their expression (Graham, A., Papalopulu, N. and Krumlauf, R., Duboule, D. and Dolle, P.; personal communications; reviewed in Refs. 19 and 53). In the Hox1 cluster, the order of the loci is 1.1, 1.2, 1.4, 1.6, and the sequence of the proteins suggest correspondence to *Antp*, *Dfd*, and *lab*. Hox1.1 and 1.2, and the most closely corresponding loci in the Hox2 complex (Hox 2.2 and 2.3), are all closely related to *Antp* (Table I). Hox1.4 and 2.6 are closely related to

each other and to *Dfd* (Table I). Hox1.6 is closely related in sequence to *lab* (Table I); a corresponding Hox2 gene has not yet been identified. If the expression patterns of Hox1 genes correspond to the Hox2 genes (in keeping with the idea that the two clusters arose through a duplication event), then the Hox1.6 gene would be expressed more anteriorly than the Hox1.4 gene, which in turn would be expressed more anteriorly than the Hox1.1 or 1.2 genes. These expression pattern would correlate with the anterior-posterior order of *Antp, Dfd* and *lab* expression, *lab* being active in the most anterior regions. Therefore, there are some hints that the gene order in the mammalian clusters may be related to the gene order of the fly clusters, both in sequence homology and in anterior-posterior expression patterns (Duboule, D. and Dolle, P.; Graham, A., Papalopulu, N. and Krumlauf, R.; personal communications). This suggests that Lewis' observation that gene order corresponds to site of function in the embryo applies not only to the *Drosophila* complexes but also to the mammalian genes, indicating a very ancient gene arrangement the function of which remains unknown. In any case, the distinct spatial and temporal expression patterns observed for many of the non-*Drosophila* homeobox-containing transcripts are consistent with a role in pattern formation in higher eukaryotes.

III. Models of homeodomain function

III-A. Helix-turn-helix proteins

The first clue about the function of the homeodomain came from the basic character of the predicted protein sequence. About 30% of the amino acids are basic, suggesting the possibility that the homeodomain associates with nucleic acid. The nuclear location of all of the homeodomain-containing proteins examined to date is also consistent with such a role. A more precise clue came from the observation [120] that the homeodomain contains a region that is similar in sequence to the α-helix-turn-α-helix sequences that are found in many bacterial DNA-binding proteins [121]. The first (more N terminal) helix is eight amino acids long, the turn is three amino acids long, and the second helix is nine amino acids long. These bacterial proteins bind as dimers to DNA sequences about 15–20 bp long, and the binding sites often have dyad symmetry. In the λ *cro* protein, for example, one helix, the 'recognition helix', lies in the major groove of the DNA and is thought to play the major role in sequence recognition. An α-helix is about 12 Å in diameter, and the major groove of B form DNA is 6–8 Å deep and about 12 Å wide. Therefore the fit is reasonable. In λ *cro*, the two recognition helices, one is each protein monomer, are 34 Å apart [122] which is the proper spacing to fit into two adjacent major grooves on the same face of the DNA double helix. The more N-terminal of the two helices lies over the recognition helix and makes contacts with phosphate groups; this helix, which we will refer to as helix 2, is believed to stabilize the interaction of the recognition helix with the DNA. Contacts of parts of the protein outside the recognition helix may also be important for sequence-specific DNA binding.

The critical residues in the helix-turn-helix framework (Table II) are a hydrophobic residue at the fourth and eighth positions in helix 2, an alanine in the fifth position in helix 2, a glycine (or sometimes cysteine or serine) in the first position of the turn followed by a hydrophobic residue, an isoleucine or valine at the fourth position of the recognition helix, and a hydrophobic residue at the seventh position of the recognition helix. All of these attributes are common to the homeodomain sequences that are listed in Table I. The helix-turn-helix part of the homeodomain is located at positions 32–51 using the coordinates of Table I. A Garnier-Robson analysis [123] predicts two α-helices in this region of the protein, in keeping with the model, and no helix-breaking residues are found in the predicted helical regions in any of the known homeodomain sequences. According to the analysis, the helices may extend beyond the lengths seen in the bacterial proteins, which would allow the most highly (i.e., absolutely) conserved homeodomain residues to be included in helix 3. Recently, NMR analysis of a purified Antennapedia homeodomain has provided direct support for a helix-turn-helix structure [123a].

An additional α-helix is predicted for the homeodomain by the Garnier-Robson analysis. We will refer to this helix, without proof of its existence, as helix 1. Helix 1 is predicted to extend from about amino acid 9 to amino acid 26. This region includes five of the most highly conserved amino acids in the homeodomain, i.e., those in positions 9, 13, 17, 21 and 26 (Table II). P. O'Farrell (personal communication) has observed that if these conserved amino acids are wound into an α-helix, they would form a hydrophobic face on one side of the helix. He has proposed that this part of the homeodomain is involved in protein–protein contact, for example to allow formation of protein dimers.

In the absence of structural studies, it is not strictly correct to consider the sequences encoded by homeoboxes to be a structural domain of a protein. However, there is some indirect evidence that the homeodomain is a functional unit, and perhaps a physical domain, of the proteins that contain it. First, homeodomains alone, tested as β-galactosidase fusion proteins, are capable of sequence-specific DNA binding (see below). Second, the homeobox is usually found as a separate exon. Third, a mutation that joins the homeobox exon of one *Drosophila* gene to the non-homeobox protein-coding exons of another produces a functional protein [124,125]. Fourth, in general, the homeodomain sequence has been much more highly conserved during evolution than the

surrounding protein sequences. A large body of data concerning prokaryotic helix-turn-helix proteins now exists. Although the relationship between homeodomains and these proteins is not proven, these data provide a useful framework for ongoing studies of homeodomain function and therefore merit further discussion.

III-B. Binding site specificities of bacterial proteins

Many bacterial proteins have been proposed to have a helix-turn-helix structure, but only five crystal structures of helix-turn-helix proteins have been reported [122,126–130]. All of these proteins are dimers. One of the structures is of the DNA-binding fragment of the λ repressor; one is of a phage 434 repressor fragment bound to its 14 bp operator. The other two are of complete proteins, the *CAP* protein and the *trp* repressor. The specificity of a helix-turn-helix protein for a particular DNA sequence is dependent upon the recognition helix sequence, as has been shown by changing specific amino acids of the recognition helix of 434 repressor [131] and of CAP [132] to cause changes in DNA binding specificity, by substituting the recognition helix of 434 cro protein for the recognition helix of 434 repressor [133], and by isolating mutations in the recognition helix of the *trp* repressor, as negative complementers, that affect the DNA binding [134]. These data support the structure–function relationship proposed for the recognition helix, and suggest that the corresponding helix in the homeodomain affects DNA-binding specificity.

Based on the crystal structures of λ *cro*, amino acids in the third position of the turn, in the first two and sixth positions of the recognition helix, and an amino acid two amino acids downstream of the recognition helix are predicted to contact bases in the major groove [135–137]. Amino acids in the corresponding positions of a homeodomain may therefore be critical for DNA sequence selection. Residues in the first and fifth positions of helix 2 and the fifth and ninth positions of the recognition helix are predicted to make contacts with the DNA backbone. Mutations that affect DNA binding by λ repressor (without detectably unfolding the protein) cluster in the helix-turn-helix region [138–140], and *trp* repressor mutations that compete for DNA binding with wild-type repressor also cluster in the corresponding region [134]. The specificity of binding of the CAP protein can be altered by any of three mutations of the second position of the recognition helix [132]. Mutations that alter the first and second positions of the *lac* repressor recognition helix cooperatively change the specificity of binding [141]. Only one point mutation in the putative helix-turn-helix region of a homeodomain has been reported. It is an alanine to valine change in the *ftz* protein in the fifth position of helix 2, i.e., in the residue predicted to contact the valine or isoleucine in the fourth position of the recognition helix [120]. The mutation causes the *ftz* protein to become temperature-sensitive in vivo and would not be predicted, based on knowledge of the bacterial proteins, to affect DNA-binding specificity.

In Table IIB, the assortment of amino acids found at each position of the homeodomain is shown. The framework positions, the hydrophobic residues in the fourth and eighth position of helix 2, in the central position of the turn, and in the seventh position of the recognition helix are all very highly conserved among the sequences: the fourth position of helix 2 is always an isoleucine, leucine, valine or methionine; the eighth position is nearly always a leucine, as is the central residue of the turn; and the seventh residue of the recognition helix is always a tryptophan. The fifth position of helix 2 (i.e., position 36 of the sequences in the tables), commonly an alanine in bacterial proteins, is usually an alanine (73/87 cases) in the homeodomains. The first residue of the turn, which is most often a glycine in bacterial proteins, is quite variable among the homeodomains, although it is a glycine in several of the cases. Four amino acids are invariant among the 83 non-yeast sequences. All of the invariant residues occur at the C-terminal part of the predicted recognition helix and just downstream of it. The other residues in, and downstream of, the recognition helix are, in general, the least variable of any part of the homeodomain.

In the positions of the homeodomains that would be expected, from the model, to control specificity of DNA binding, there is considerably more variation than there is among the framework amino acids. The most variable positions of the recognition helix, the first, second, fifth, sixth and ninth positions, are exactly the amino acids predicted from the bacterial models to contact DNA. The amino acids two residues downstream of the recognition helix, which is also predicted to contact DNA, is also somewhat variable and is flanked by absolutely invariant amino acids. The homeodomains that are most divergent in the recognition helix are the *C. elegans* homeodomains encoded by the *mec3* and *unc-86* genes, and the mammalian homeodomains in the *unc-86* class. The *Drosophila prd* segmentation gene is closely related to the *Drosophila gsb* region genes BSH4 and BSH9 (also segmentation genes) by DNA cross-hybridization; they are also noticeably related in their recognition helix sequences, by having a threonine at the end of helix 2, and by having arginines in positions 56 and 58. Homeodomains of the *en* class stand out by having lysine in position 53 and phenylalanine in position 9. Homeodomains of the *unc-86* class stand out by having a cysteine in position 51 in the recognition helix. Thus the identities of generally highly conserved residues make it possible to classify some of the homeodomains into groups, just as overall sequence similarity can. The

consensus sequences for the different classes of homeodomains are shown in Table IIA. Undoubtedly the number of classes will increase and become clearer as more homeodomain sequences are obtained.

The original hypothesis that homeodomains contain helix-turn-helix structures was based on three homeodomain sequences [120]. The addition of more than 80 additional sequences does not reveal any major contradictions of the hypothesis but instead provides additional support for the idea.

III-C. DNA-binding studies with homeodomains

The sequence similarity of homeodomains to known DNA-binding proteins is strong enough to be suggestive, but not so strong as to provide certainty as to the function(s) of the proteins. More direct evidence for DNA binding has been provided by studies in which *Drosophila* proteins were tested for their abilities to preferentially bind to specific sequences of DNA. These studies are made more difficult by the complete absence of knowledge about which, if any, sequences are bound in vivo, or even which genes are directly regulated by any homeodomain-containing gene. A first step is just to show that the proteins have affinity for specific DNA sequences without immediately approaching the more difficult problem of finding functionally important binding events. This step was first taken by Desplan et al. [142] who showed that the protein product of the *engrailed* locus, when made in bacteria as a β-galactosidase fusion protein, is capable of sequence-specific DNA binding. They also showed that the *en* homeodomain, with only a small amount of surrounding *en* protein, is capable of the observed DNA binding. Studies with mammalian homeodomains also demonstrated sequence-specific binding [143,144]. Due to the types of protein used, it was not possible to measure the affinities of the proteins for the binding sites quantitatively, nor were the exact limits of the binding sites defined.

Recent experiments with fly proteins made in bacteria have employed DNase I footprinting studies to better define the exact sequences of the binding sites. Hoey and Levine [145] have used four different homeodomain-containing proteins, all produced as unfused full length proteins in bacteria using a T7 expression system, to map binding sites. The proteins were from the *Drosophila eve, en, prd* and *zen* genes. The first three are segmentation genes, the fourth is a gene involved in dorsal/ventral differentiation. The four proteins have very different homeodomains (Table I). In particular the homeodomains differ in the recognition helix in ways that lead to the prediction that they will bind to different sequences.

In the binding experiments, the DNA used was from the *eve* and *en* genes. The experiments employed crude extracts of bacteria, and the protein extracts went through a 4 M guanidine step which means the protein was denatured and renatured. Therefore some of the differences observed between the behaviors of the four proteins may have been due to different protein stabilities or to protein folding problems. *eve* protein was found to bind specifically to sequences near the 5′ end of each of the two genes. The binding of *eve* to *en* DNA could have biological significance because *eve* is known to be an activator (directly or indirectly) of *en* [66,67]. Based on footprint data, the consensus sequence for the binding sites is TCAATTAAAT, a sequence that is found in up to three copies at some of the bound sites. However, some of the bound sites do not contain a sequence similar to this consensus and instead are GC-rich. *eve* protein binds to both types of site with similar affinity. *zen* protein binds to the GC-rich sites but *en* and *prd* proteins bind only weakly to them. None of these proteins binds as strongly to the GC-rich sites as does *eve* protein. None of the binding sites has yet been tested for function in vivo.

The sequences to which *en* protein binds have also been determined using DNase I footprinting. The consensus sequence for binding is 5′TCAATTAAAT 3′, which is identical to the consensus sequence shown above [146]. The *ftz* protein also binds to this sequence [146], as does the *Xenopus* homeodomain-containing protein x1hbox1 [147]. Desplan et al. [146] also showed that single point mutations in the consensus sequence could drastically reduce the binding affinity of the site, and that both *en* and *ftz* fusion proteins also bind to a repeated TAA sequence that has been found to be a binding site for *Ubx* protein. The *ftz* homeodomain bound about equally well to the TCAATTAAAT and TAA repeat, while the *en* homeodomain had a strong preference for the TCAATTAAAT sequence.

The protein product of the *Ultrabithorax* (*Ubx*) homeotic gene has been found to bind both to specific DNA sequences with the consensus sequence 5′TAATAATAATAATAA 3′ and to repeats of the hexanucleotide TAATCG [148,149]. Other variations on these sequences are bound as well (Beachy, P., personal communication). The bound sites are 40–90 bp long, based on footprint analyses. The *Ubx* protein used was made in bacteria but was not joined to β-galactosidase and was not denatured prior to DNA binding. Using electron microscopy, it has been shown that the *Ubx* protein can bind to two sites on a DNA fragment and by doing so curl the fragment into a loop (Beachy, P., personal communication). This was demonstrated using two sites about 235 bp apart on a DNA fragment corresponding to the *Ubx* RNA 5′leader region.

In studies with β-galactosidase protein fused to *ftz* or *Deformed* (*Dfd*) (a homeotic gene) homeodomains, 25 different footprint sites were identified in a scan of DNA from the *Antp* gene [151]. The proteins bound to specific DNA sequences but no single consensus se-

quence was found. Instead three consensus sequences could be derived from the data (Laughon, A., personal communication). One, 5′TTT(A/C)(T/C)NTTAATTGCTT(T/A)(T/A)AT3′, with TTAATTGC being the most conserved 'core', is similar to the sequence found in the four protein study described above and to the sequence bound by *en* and *ftz* homeodomain fusion proteins in the study by Desplan et al. [146]. The second, 5′TATTTAATAATAATGNNATNA3′, is similar to the sequence bound by *Ubx* protein [149]. The third is 5′GTAATCGTA3′. Each consensus sequence is based on at least eight footprinted sites. The three types of site have in common the TAAT sequence, which therefore stands out as possibly a key attribute of binding sites for these proteins. A purified *Antp* homeodomain also binds to TAAT [150]. The other consensus sequences shown above also contain TAAT or its complement, as does the operator site for the yeast homeodomain protein MATα2 [152].

Thus, homeodomains are capable of sequence-specific DNA binding in vitro. Do those interactions occur in vivo and affect the transcription of other genes? Recent experiments directed at this problem using cultured *Drosophila* cells and a transient expression system have shown that the *Ubx* protein can repress transcription from an *Antp* promoter and stimulate transcription from the *Ubx* promoter [152a], while *Antp* and *ftz* proteins can activate transcription from the *Ubx* promoter [152b]. Sequences near the *Ubx* promoter that respond to *Antp* activation map to the RNA leader region and may be similar or identical to the sequences there that are bound by *Ubx* protein. *Antp* protein tested in vitro also binds to DNA corresponding to the *Ubx* RNA leader (Hayashi and Scott, unpublished data).

A transcriptional activator found in the rat pituitary gland has been found to contain a homeodomain. This 'Pit-1' protein, which is quite probably the same as the protein GHF-1, was purified on the basis of its binding to an important *cis*-acting regulatory sequence of prolactin and growth hormone genes [56,62,153]. Pit-1 protein may have roles in addition to its presently known ones. The sequence recognized by the Pit-1 protein is 5′(T/A)(T/A)TATNCAT3′, which could be related to the TAAT theme found in the experiments described above. Pit-1 is in a divergent class of homeodomains with a recognition helix sequence quite different from the proteins used in the *Drosophila* DNA binding experiments and therefore would be expected to have a different preferred binding site.

The other mammalian transcription factors, the OTF or OCT proteins, have been found to contain homeodomains [57–62] and to be in the same POU class as Pit-1 (Tables I and II). The factors bind to the sequence ATGCAAAT. This sequence is an important component of the *cis*-acting regulatory sequences of H2B histone genes [154] and of immunoglobulin promoters and enhancers [155,156,160]. The OCT-1 factor, which is probably identical to OTF-1 and has also been called OBP100, is found in many cell types [157,158], whereas OTF-2 (probably the same as OCT-2) is B-cell-specific and therefore is likely to control the B-cell specificity of immunoglobulin gene expression [159–163]. OTF-1 has been found to be the same as NF-III, a DNA replication factor identified in an in vitro Adenovirus DNA replication system [163]. NF-III/OTF-1 was found to activate DNA replication, which now raises the interesting possibility that a homeodomain protein could be involved in controlling DNA replication as well as transcription. Pit-1 and the OCT factors provide clear cases of mammalian transcriptional regulation by homeodomains.

The DNA binding and transcription studies have shown that the proteins are capable of sequence recognition and transcriptional regulation and therefore the helix-turn-helix hypothesis may well be correct. The potential problems with in vitro DNA binding studies are clear: the proteins produced in bacteria may have properties different from the normal proteins, other proteins involved in the functions of the homeodomain proteins may be missing from the assays, and small cofactors that alter binding properties (such as tryptophan in the case of *trp* repressor, Ref. 130) may be missing or present in the wrong concentrations. Apart from these problems there remains considerable uncertainty about the range of sequences that can be bound by homeodomain proteins. An obvious and only partly answered question is how, given the similarity of the homeodomains and their DNA-binding behavior, the proteins control specific aspects of development. The experimental results to date appear to demonstrate that many of the proteins can bind to similar sites, and that at least some of the proteins can bind to more than one type of site. These issues are particularly relevant because it is clear that multiple *Drosophila* homeodomain-containing proteins can coexist within the same cell (e.g., Ref. 165). There also appear to be cases in mammalian homeobox expression where multiple genes are active in the same cells (reviewed in Ref. 53), but the in situ hybridization analyses used have insufficient resolution to be certain of this result. If, as in the case of the bacterial helix-turn-helix proteins and yeast MATα2, homeodomain proteins are multimeric, there would opportunities for the formation of heterogeneous multimers. Heterodimers might, for example, have DNA binding or other properties different from those of the related homodimeric proteins. For all of these reasons it is now crucial to use functional tests to establish which sites are used in vivo and to determine whether proteins compete for sites or cooperative in bindin

IV. Genetic analyses of homeodomain function

IV-A. Regulatory interactions among Drosophila homeobox-containing genes

In much the same way as studies of prokaryotic DNA binding proteins and yeast homeodomains suggest models of *Drosophila* homeodomain function, the analysis of the network of genes controlling *Drosophila* development has implications for the study of homeobox genes in higher eukaryotes. Many of the *Drosophila* homeobox-containing genes are part of the developmental network. In general, there does appear to be a good correlation between the regions of expression and function for the *Drosophila* homeobox genes, many of which are expressed in intricate and dynamic spatial and temporal patterns during development. Thus, considerable work has been directed to the study of how homeobox gene expression is regulated. While genetic studies have revealed much about the developmental role of homeobox genes, considerably less is known about the molecular mechanisms underlying their function.

Since homeobox genes are thought to act as transcription regulators of other loci, the identification of the targets of homeodomain function has also represented a major goal. In *Drosophila*, a small number of targets of homeodomain function have been identified. In addition, genes which act to regulate homeobox gene expression have also been found. Rather than review the functions of the *Drosophila* homeobox genes in detail, which has been done elsewhere [5,8,9,18,19], we will discuss cases of regulatory interactions involving homeodomain proteins and/or homeobox genes. Interactions among the genes have generally been studied by looking at the expression of one of the genes in embryos mutant for another. This sort of experiment establishes a hierarchy between the genes, but cannot say whether the interaction is direct or indirect, an issue that is particularly important in cases where the product of one gene is hypothesized to regulate a target gene by binding to it. Among the various interactions between regulatory genes, several stand out as possible cases of direct effects of the product of one gene on the transcription of another. As discussed in the previous section, the identification of the targets of homeodomain-containing proteins will greatly aid the analysis of DNA binding by homeodomains.

At the earliest stages of zygotic development, expression of the segmentation genes depends on maternally provided RNA and the corresponding proteins. Some of the earliest genes to become active are the 'gap' genes that are expressed, and function, in broad regions of the embryo. One of these, *hunchback* (*hb*), encodes a zinc finger protein and is expressed in the anterior region of the embryo [166]. The activation of the *hb* gene has been shown to depend on the activity of a maternally provided homeodomain protein encoded by the gene *bicoid* (*bcd*) [167]. The responsive *hb* DNA has been delimited to a 300 bp region just upstream of the transcription initiation site and sites through which *bcd* protein can activate *hb* transcription have been identified [167a].

After the earliest patterns of zygotic gene activity are established the segmentation genes interact so that their initially broad transcription patterns are refined into specific patterns such as transverse stripes. The first homeobox-containing genes to become active (except for maternally encoded proteins) are the 'pair-rule' genes that are characterized by mutations that cause embryos to develop with about half the normal number of body segments. At least three of the genes in this class, *fushi tarazu* (*ftz* [17,16]), *even-skipped* (*eve* [66,67]), and *paired* (*prd* [168]), contain homeoboxes. *ftz* stripes form in the absence of *eve* function but they are abnormal in spacing, the first stripe is often missing, and the stripes disappear prematurely [67,169]. Therefore *eve* function appears to be required for proper formation and maintenance of the transcription of *ftz* in stripes; the *eve* protein could act directly on *ftz*. Mutations in *prd* affect neither *ftz* [169] nor *eve* [170] expression, and *ftz* mutations do not affect *eve* expression in the blastoderm embryo. Curiously, *eve* function is required for proper *ftz* expression in the developing nervous system [171], so the cellular context appears to be able to affect the interactions that occur.

After the pair-rule genes have come to be expressed in transverse stripes, another class of genes becomes active: the segment polarity genes. These genes are also expressed in transverse stripes but the stripes are narrower and occur at a density of one per segment primordium rather than one for each two segment primordia as is the case for pair-rule genes. The best studied segment polarity gene is *engrailed* (*en*), which contains a homeobox. The *en* stripes are controlled by several of the pair-rule genes, including the homeobox-containing genes *ftz*, *eve*, and *prd* [66,67,172–175]. These three genes are all activators of *en* stripes. *ftz* and *eve* have been shown to be negative regulators of another segment polarity gene, *wingless* [176]. The interactions among segmentation genes therefore lead to three important conclusions: (i) a gene can be a positive regulator of one target gene and a negative regulator of another; (ii) the cellular context may affect which gene interactions occur; and (iii) there are several cases in which the timing of expression of interacting genes suggests that the interaction could be direct, the product of one gene binding to the other gene.

The initial pattern of homeotic gene expression in the embryo appears to be controlled by various segmentation genes including *ftz*. Each of the three homeotic genes *Ubx*, *Antp*, and *Scr* is transcribed at the blastoderm stage in a discrete region that overlaps with one of

the transverse stripes where *ftz* protein accumulates. In the absence of *ftz* function, the peaks of expression of the three homeotic genes in the blastoderm stage embryo do not form [177]. Therefore the *ftz* protein may be a direct activator of transcription of each of the three homeotic genes. *en* negatively regulates *Ubx* in the posterior regions of abdominal segments [175], so the *en* protein homeodomain could be a direct repressor of *Ubx* transcription.

After the initial patterns of homeotic gene transcription are set up, there are continuous changes and refinements of the patterns. Some of these changes, as well as the maintenance of some aspects of the patterns that do not change, require interactions between homeotic genes. For example, the genes of the bithorax complex, *Ubx*, *abdA* and *AbdB*, negatively regulate *Antp* expression in posterior regions of the embryo [178–180], and *abdA* and *AbdB* negatively regulate *Ubx* expression in the abdominal regions of the embryo [180]. These interactions could be direct, the homeodomain-containing products of one gene binding to another gene to lower (not eliminate) its transcription. This molecular model of the interactions has not yet been tested directly. In one case a *cis*-acting element that responds to homeodomain proteins has been roughly mapped. The *Antp* gene has two promoters. The *cis*-acting elements of one of them, P2, has been studied in vivo using *lac Z* gene fusions and P-element-mediated introduction of the engineered genes into the germ line [182]. The P2-*lac Z* fusions are negatively regulated by *Ubx*, *abdA*, and *AbdB*, as P2 normally is. The region responsive to *Ubx* and *abdA* was mapped to a region extending from the promoter to 2 kb upstream. Additional experiments will allow a more precise mapping of the control elements.

IV-B. The functions of nematode homeobox genes

Three *Caenorhabditis elegans* genes have been discovered for their effects on development and subsequently found to contain homeoboxes. The first, *mec-3*, instructs cells to develop into 'touch cells' (mechanosensory neurons) rather than follow another developmental pathway [28]. This pathway choice is conceptually similar to the cell fate decisions controlled by *Drosophila* homeotic genes, and therefore *mec-3* is a kind of homeotic gene. A decision between alternative neural fates is also controlled by the *cut* gene of *Drosophila* [183] and *cut* also contains a homeobox [38]. Both of these genes therefore fall into a broadened category of homeotic genes: those that control cell fates as opposed to body segment fates.

The second *C. elegans* gene with a homeobox is the *mab-5* gene [42,216]. *mab-5* controls the development of an array of cells, from different tissue types and lineages, that form structures characteristic of the posterior of the worm. In general without *mab-5* function, the posterior cells develop in anterior patterns; some of the transformations alter male-type development to hermaphrodite pathways. Therefore the phenotype of *mab-5* mutations is conceptually very similar to homeotic phenotypes in *Drosophila*. The third *C. elegans* gene found to have a homeobox is the *unc-86* gene [43] which controls cell lineages [44]. In the absence of *unc-86* function, cell lineage patterns are reiterated rather than diverging as they do in wild-type worms. The affected cells act in quite different parts of the worm. The changes in cell lineages indicate a failure of cell to differentiate.

A fourth homeobox-containing gene (JML1001) has been found in the *C. elegans* genome (Hawkins, N. and McGhee, J., personal communication), but nothing is yet known about its function or expression. The three homeoboxes whose functions are known provide further evidence for the association of homeoboxes with genes that control development.

IV-C. Tests of function in Xenopus

In *Drosophila*, two kinds of mutation are often found in homeobox genes: mutations that inactivate the gene and prevent it from performing its normal function in the cells where it is needed, and mutations that result in expression of the gene in cells where it is normally inactive. It is currently very difficult to obtain any mutations in vertebrate genes that are identified as molecular clones. To circumvent this problem, approaches have been developed to mimic the phenotypes expected from two types of mutations: dominant gain-of-function and recessive loss-of-function mutations. An experiment of this sort has been done by injecting synthetic RNA made from the cloned *Xhox-1a* gene into *Xenopus* eggs [184]. RNA encoding a truncated form of the protein has no effect, but RNA encoding the full protein causes dramatic developmental defects. The most striking effect is a disruption of the somites; the metameric pattern is almost completely lost. The normal pattern of expression of *Xhox-1a* is not known in detail, but it is absent in the head and present both dorsally and ventrally in the trunk. Based on dissections and RNA blots, the RNA appears to be predominantly in the somitic mesoderm. Therefore the injection experiment may cause the protein to act in somitic mesoderm cells where (or when) it is normally inactive. If somitic metameres are regarded as analogous to segments, there is a tantalizing similarity of the injection phenotype to segmentation gene mutant phenotypes in flies. However, there is good reason to question such an analogy [185]. The safer, and probably better, analogy is that pattern formation is affected in both cases. This is the first evidence for a function in pattern formation of a vertebrate homeobox gene, although it is a demonstration of what the gene product can do to interfere with

proper development rather than a demonstration of its normal developmental function. Nonetheless, the experiment is a significant step forward. This approach has also been taken to analyze the developmental role of Xhox3 (Ruiz i Altaba, A. and Melton, D.A., unpublished data). Xhox3 mRNA is normally found in highest concentration in the posterior pole of the anterior–posterior axis of the mesoderm. Interestingly, microinjection of Xhox3 mRNA into *Xenopus* embryos was found to cause defects in, or deletions of, anterior structures, suggesting that Xhox3 does indeed play a role in the establishment of maintenance of anterior–posterior identities in the developing *Xenopus* embryo.

A different and complementary approach recently has been taken in the study of x1hbox1 function [147]. x1hbox1 activity was blocked by microinjection of antibodies directed against this protein into developing *Xenopus* embryos. The microinjection of such antibodies might mimic the phenotype expected from loss of function mutations in this gene. The antibodies were found to diffuse throughout the injected embryos and were stable for at least 48 h post injection. A reproducible (though low frequency) consequence of injection is the deletion of dorsal fin structures. Though the affected structures are derived from regions that normally express x1hbox1 in the neural crest, other regions that normally express the protein are not affected by antibody microinjections. Nonetheless, this technique represents a potentially powerful approach for dissecting the function of homeodomain-containing proteins in organisms not amenable to genetic analysis. Another potentially very powerful approach to homeobox gene function in vertebrates is the use of gene disruption in transgenic mice (e.g., Ref. 186), but no experiments of this sort involving homeoboxes have been reported yet.

IV-D. The MATα2 repressor: a yeast paradigm for homeodomain functional studies

Because yeast is a single celled organism, it is not possible for yeast homeobox genes to control pattern formation. However, they could play a role in cell differentiation and in at least one case they do. Given the relative ease and rapidity with which yeast genetic analysis can be conducted, a great deal of progress has been made in the study of yeast homeodomain function. These studies have important implications for the study of homeodomain function in higher organisms. The first relationship between the homeobox genes of *Drosophila* and a yeast gene to be noted was the case of the yeast MAT locus, which controls the mating type of the yeast cell (reviewed in Refs. 45 and 47). The MAT locus produces different proteins depending which of two alternative 'cassettes' of DNA has been brought to the MAT locus from donor loci by gene conversion. A yeast cell is directed to be an α mating type cell if an α cassette is present at the MAT locus and to be an a mating type cell if an a cassette is present at the MAT locus. The protein coding capacity of MAT depends on which cassette is present. The α allele of MAT encodes two proteins; one is a protein called α2 that has been shown to be a repressor of a-specific genes [152,189]. The a allele of MAT encodes a protein called a1 that has been shown to repress, in collaboration with α2, haploid-specific genes [188,190,191]. Both α2 and a1 have sequence similarity to the homeodomain (Table I; Refs. 120 and 192). The region of greatest similarity is in the putative helix-turn-helix of the homeodomain. Nothing is yet known about the structure of the yeast MAT proteins, but a great deal has been learned about the function of α2, and the information may serve to guide studies of other homeodomain-like sequences.

The MATα2 protein is being intensively studied as a case of a eukaryotic repressor. This protein turns off transcription of a set of target genes that are normally only expressed in *a* mating-type yeast cells [152]. The protein was the first yeast protein to be described as having a homeodomain, in part because its sequence similarity to the *Drosophila* homeodomains extends beyond the putative helix-turn-helix region. MATα2 protein has been overexpressed in *E. coli* and purified but its three-dimensional structure is not yet known. The protein is a covalently (disulfide) linked dimer with identical subunits [193]. The homeodomain part of the protein has been shown to be sufficient for sequence-specific DNA binding, but this part of the protein alone is not sufficient for repression in vivo [194]. Therefore as with some bacterial protein, one domain of the protein may be involved in DNA binding and another in interactions with other factors to repress transcription. The second MAT product with homeodomain homology, α1, also has been shown to bind to specific sequences near genes that it regulates [195]. In contrast to α2 protein, the α1 protein is an activator of transcription of α-specific genes.

The binding of MATα2 protein to the 32 bp operator upon which it acts occurs through two contact points about two and a half turns of the DNA helix apart [193]. Contact with the center of the operator has not been detected, despite the obvious sequence conservation of the centers of operators from different genes regulated by MATα2. This paradoxical result is now understood because another protein, called GRM, binds to the center of the operator [196]. GRM interacts with MATα2 protein through protein–protein contact involving the N-terminal domain of MATα2. GRM is found in yeast cells of all three mating-types and therefore probably has more functions than just to interact with MATα2. GRM may be identical to a transcriptional *activator* protein called PRTF [195] that works together with the second protein encoded by the MATα

locus, the α1 product, which is required to turn on α-specific genes.

In addition to its interaction with GRM, MATα2 protein has been found to interact with the other MAT product that contains a homeodomain, **a1** protein [197]. Both the α2 and **a1** proteins are present in diploid yeast cells, and in such cells a set of genes called the haploid-specific genes is repressed [188]. In such cells it appears that α2 protein dimers turn off transcription of *a*-specific genes and α2 and **a1** proteins acting together turn off transcription of haploid-specific genes. The DNA binding specificity of α2 protein is changed in the presence of **a1** protein, as had been suggested by the finding [189,198] that a 29 bp consensus sequence at haploid-specific genes differs from the operator acted upon by α2 alone. The spacing of the contact points for α2 in α2 operator sites is different from the spacing in **a1**/α2 sites, suggesting that the α2 protein is binding differently to the two types of sites, perhaps because **a1** protein shares in the binding or because **a1** protein alters the conformation of α2 protein.

These findings with MAT products may provide some idea of what can be expected with homeodomain proteins in higher eukaryotes: interactions with non-cell-type-specific proteins and with other homeodomain-containing proteins, and binding to a variety of different DNA sequences. Thus, the results obtained in DNA-binding studies using purified homeodomain-containing proteins should be cautiously interpreted. As is described above, the data on DNA binding by higher eukaryotic homeodomains do provide some indications of such interesting phenomena.

V. Conclusions

It seems clear now that at least one function of homeodomain proteins is to regulate transcription by binding to specific DNA sequences, as has been shown most clearly for the Pit-1 and OCT cases. It is still possible, however, that the proteins, or some of them, have other functions instead or as well. The relationship of the homeodomain to the helix-turn-helix structure remains to be proven, but the data to this point are consistent with a remarkably direct extrapolation from studies of transcriptional regulation in bacteria to the gene networks that control development. It is also surprising that so many of the *Drosophila* regulatory genes, nearly all of which were first found genetically, contain homeoboxes. Together with zinc finger proteins, the homeodomain proteins account for a very large fraction of the *Drosophila* regulatory proteins whose sequences have been determined. One might have expected a greater diversity in the composition of regulatory proteins even within the group of segmentation and homeotic genes. The high correlation of homeoboxes with regulators of development in flies is reason for optimism that the homeoboxes in other organisms will succeed in leading us to regulatory genes, and the first experiments with manipulating frog homeobox genes add to the optimism.

Major hurdles that lie ahead are to understand the specificity of homeodomain interactions with target genes and with the transcription apparatus. The functions of parts of the proteins outside of the homeodomain need to be understood. The interactions of homeodomain proteins with each other, in competition or cooperation, and with other classes of regulatory proteins, are very likely to be crucial to the regulatory systems. Although we can now point at cases where the product of gene A is likely to act directly on gene B, there is only a very fuzzy picture, even in flies, of how the whole regulatory circuitry is integrated. As precise mechanisms of homeodomain protein functions are worked out, we can expect additional clues as to how the systems that control multicellular development evolved. The homeodomain is a powerful 61 amino acid unit, and the reasons for its usefulness to organisms and to experimentalists will become clearer as this exciting field moves forward.

Acknowledgments

We are grateful to the many members of the rapidly growing homeobox research community who contributed unpublished information and very helpful ideas to the review. The amount of information has grown extremely fast, and we apologize for any errors in interpretation, nomenclature, or attribution – we have spent a great deal of time to try to catch any such problems, but no doubt some remain. We also thank Cathy Inouye for her patience and speed in preparing many drafts of the manuscript, and Dr. Susan Dutcher, Dr. Gary Stormo and John Bermingham for critical readings of the work. We are grateful to Shelly Greenfield for assistance in the proofreading of the homeodomain sequences. Among the computer programs we used, the EuGene sequence analysis package distributed by the Molecular Biology Information Resource, Department of Cell Biology, Baylor College of Medicine and the Phylip program from Joseph Felsenstein at the University of Washington were especially helpful. The research in our laboratory is supported by grants from the American Cancer Society, the March of Dimes, the National Institues of Health, and the Searle Scholar Program. J.W.T. gratefully acknowledges a postdoctoral fellowship from the Jane Coffin Childs foundation.

References

1 Bateson, W. (1894) Materials for the Study of Variation, MacMillan and Co., London.
2 Slack, J.M.W. (1985) J. Theor. Biol. 114, 463–490.
3 Lewis, E.B. (1978) Nature 276, 565–570.

4 Kaufman, T.C., Lewis, R. and Wakimoto, B. (1980) Genetics 94, 115–133.
5 Akam, M. (1987) Development 101, 1–22.
6 Duncan, I. (1987) Annu. Rev. Genet. 21, 285–319.
7 Peifer, M., Karch, F. and Bender, W. (1987) Genes Dev. 1, 891–898.
8 Scott, M.P. and Carroll, S.B. (1987) Cell 51, 689–698.
9 Ingham, P.W. (1988) Nature 335, 25–34.
10 Bender, W., Akam, M.A., Beachy, P.A., Karch, F., Peifer, M., Lewis, E.B. and Hogness, D.S. (1983) Science 221, 23–29.
11 Garber, R.L., Kuroiwa, A. and Gehring, W.J. (1983) EMBO J. 2, 2027–2034.
12 Scott, M.P., Weiner, A.J., Polisky, B.A., Hazelrigg, T.I., Pirrotta, V., Scalenghe, F. and Kaufman, T.C. (1983) Cell 35, 763–776.
13 Karch, F., Weiffenbach, B., Pfeifer, M., Bender, W., Duncan, I., Celniker, S., Crosby, M and Lewis, E.B. (1985) Cell 43, 81–96.
14 Pultz, M.A., Diederich, R.J., Cribbs, D.L. and Kaufman, T.C. (1988) Genes Dev. 2, 901–920.
15 McGinnis, W., Levine, M., Hafen, E., Kuroiwa, A. and Gehring, W.J. (1984) Nature 308, 478–433.
16 McGinnis, W., Garber, R.L., Wirz, J., Kuroiwa, A. and Gehring, W.J. (1984) Cell 37, 403–408.
17 Scott, M.P. and Weiner, A.J. (1984) Proc. Natl. Acad. Sci. USA 81, 4115–4119.
18 Gehring, W.J. and Hiromi, Y. (1986) Annu. Rev. Genet. 20, 147–173.
19 Gehring, W.J. (1987) Science 236, 1245–1252.
20 Mavilio, F., Simeone, A., Giampaolo, A., Faiella, A., Zappavigna, V., Acampora, D., Poiana, G., Russo, G., Peschle, C. and Boncinelli, E. (1986) Nature 324, 664–668.
21 Krumlauf, R., Holland, P.W.H., McVey, J.H. and Hogan, B.L.M. (1987) Development 99, 603–617.
22 Regulski, M., McGinnis, N., Chadwick, R. and McGinnis, W. (1987) EMBO J. 6, 767–777.
23 Wilde, C.B. and Akam, M. (1987) EMBO J. 6, 1393–1401.
24 Wakimoto, B.T. and Kaufman, T.C. (1981) Dev. Biol. 81, 51–64.
25 Müller, M.M., Carrasco, A.E. and DeRobertis, E.M. (1984) Cell 39, 157–162.
26 McGinnis, W. (1985) Cold Spring Harbor Symp. Quant. Biol. 50, 263–270.
27 Holland, P.W.H. and Hogan, B.L.M. (1986) Nature 321, 251–253.
28 Way, J.C. and Chalfie, M. (1988) Cell 54, 5–16.
29 Carrasco, A.E., McGinnis, W., Gehring, W.J. and DeRobertis, E.M. (1984) Cell 37, 409–414.
30 Poole, S.J., Kauvar, L., Drees, B. and Kornberg, T. (1985) Cell 40, 37–43.
31 Coleman, K.G., Poole, S.J., Weir, M.P., Soeller, W.C. and Kornberg, T. (1987) Genes Dev. 1, 19–28.
32 Joyner, A.L., Kornberg, T., Coleman, K.G., Cox, D.R. and Martin, G.R. (1985) Cell 43, 29–37.
33 Joyner, A.L. and Martin, G.R. (1987) Genes Dev. 1, 29–38.
34 Dolecki, G.J. and Humphreys, T. (1988) Gene 64, 21–31.
35 Bopp, D., Burri, M., Baumgartner, S., Frigerio, G. and Noll, M. (1986) Cell 47, 1033–1040.
36 Herr, W., Sturm, R.A., Clerc, R.G., Corcoran, L.M., Baltimore, D., Sharp, P.A., Ingraham, H.A., Rosenfeld, M.G., Finney, M., Ruvkun, G. and Horvitz, H.R. (1988) Genes Dev. 2, 1513–1516.
37 Barad, M., Jack, T., Chadwick, R. and McGinnis, W. (1988) EMBO J. 7, 2151–2161.
38 Blochlinger, K., Bodmer, R., Jack, J., Jan, L.Y. and Jan, Y.N. (1988) Nature 333, 629–635.
39 Bridges, C. and Morgan, T.H. (1923) Carnegie-Inst. Wash. Publ. 327, 93.
40 Tomlinson, A., Kimmel, B.E. and Rubin, G.M. (1988) Cell 55, 771–784.
41 Saint, R., Kalionis, B., Lockett, T.J. and Elizur, A. (1988) Nature 334, 151–154.
42 Kenyon, C. (1986) Cell 46, 477–487.
43 Finney, M., Ruvkun, G. and Horvitz, H.R. (1988) Cell 55, 757–769.
44 Chalfie, M., Horvitz, H.R. and Sulston, J.E. (1981) Cell 24, 59–69.
45 Nasmyth, K. and Shore, D. (1987) Science 237, 1162.
46 Kelly, M., Burke, J., Smith, M., Klar, A. and Beach, D. (1988) EMBO J. 7, 1537–1547.
47 Oshima, Y. (1982) In The Molecular Biology of the Yeast Saccharomyces. Metabolism and Gene Expression (Strathern, J.N., Jones, E.W. and Broach, J.R., eds.), pp. 159–180, Cold Spring Harbor Laboratory, New York.
48 Arndt, K.T., Styles, C. and Fink, G.R. (1987) Science 237, 874–880.
49 Manley, J.L. and Levine, M.S. (1985) Cell 43, 1–2.
50 Colberg-Poley, A.M., Voss, S.D. and Gruss, P. (1987) In Oxford Surveys on Eukaryotic Genes (Maclean, N., ed.), pp. 92–115, Vol. 4. Oxford University Press.
51 Gruss, P., Dony, C., Föhring, B. and Kessel, M. (1987) Structure and Developmental Expression of Murine Homeobox Genes. In Molecular Mechanisms in Cellular Growth and Differentiation (Bellué, A.R. and Vogel, H.J., eds.), Proc. P&S Biomedical Sciences Symp., in press.
52 Fibi, M., Kessel, M. and Gruss, P. (1988a) Murine Hox genes – a multigene family. Curr. Topics Immun., in press.
53 Holland, P.W.H. and Hogan, B.L.M. (1988) Genes Dev. 2, 773–782.
54 De Robertis, E.M., Burglin, T.R., Fritz, A., Wright, C.V.E., Jegalian, B., Schnegelsberg, P., Bittner, D., Morita, E., Oliver, G. and Cho, K.W Y. (1988) in DNA–Protein Interactions in Transcription (Gralla, J., ed.), Vol. 95, UCLA Symposium.
55 Duboule, D., Galliot, B., Baron, A. and Featherstone, M.S. (1989) Murine homeo-genes: Some aspects of their organisation and structure. In Cell to Cell Signal in Mammalian Development (De Laat, S., Bluemask, J.G. and Nummery, C.L., eds.), NATO ASI Series, Springer Verlag, in press.
56 Nelson, C., Albers, V.R., Elsholtz, H.P., Lu, L.I.-W. and Rosenfeld, M.G. (1988) Science 239, 1400–1405.
57 Ko, H.-S., Fast, P., McBride, W. and Staudt, L.M. (1988) A human protein specific for the immunoglobulin octamer DNA motif contains a functional homeobox domain. Cell 55, 135–144.
58 Clerc, R.G., Corcoran, L.M., LeBowitz, J.H., Baltimore, D. and Sharp, P.A. (1988) Genes Dev. 2, 1570–1581.
59 Müller, M.M., Ruppert, S., Schaffner, W. and Matthias, P. (1988) Nature 336, 544–551.
60 Scheidereit, C., Cromlish, J.A., Gerster, T. Kawakami, K., Balmaceda, C.-G., Currie, R.A. and Roeder, R.G. (1988) Nature 336, 551–598.
60a Sturm, R.A., Das, G. and Herr, W. (1988) Genes Dev. 2, 1582–1599.
61 Bodner, M. and Karin, M. (1987) Cell 50, 267–275.
62 Ingraham, H.A., Chen, R., Mangalam, H.J., Elsholtz, H.P., Flynn, S.E., Lin, C.R., Simmons, D.M., Swanson, L. and Rosenfeld, M.G. (1988) Cell 55, 519–529.
63 Mlodzik, M., Fjose, A. and Gehring, W.J. (1985) EMBO J. 4, 2961–2969.
64 Levine, M., Hardin, K., Wedeen, C., Doyle, H., Hoey, T. and Radomska, H. (1985) Cold Spring Harb. Symp. Quant. Biol. 50, 209–222.
65 Macdonald, P.M. and Struhl, G. (1986) Nature 324, 537–545.
66 Harding, K., Rushlow, C., Doyle, H.J., Hoey, T. and Levine, M. (1986) Science 233, 953–959.
67 Macdonald, P.M., Ingham, P. and Struhl, G. (1986) Cell 47, 721–734.
68 Fjose, A., McGinnis, W.J. and Gehring, W.J. (1985) Nature 313, 284–289.

69 Regulski, M., Harding, K., Kostriken, R., Karch, F., Levine, M. and McGinnis, W. (1985) Cell 43, 71–80.
70 Doyle, H.J., Harding, K., Hoey, T. and Levine, M. (1986) Nature 323, 76–79.
71 Rushlow, C., Doyle, H., Hoey, T. and Levine, M. (1987) Genes Dev. 1, 1268–1279.
72 Hoey, T., Doyle, H.J., Harding, K., Wedeen, C. and Levine, M. (1986) Proc. Natl. Acad. Sci. USA 83, 4809–4813.
73 Mlodzik, M., Fjose, A. and Gehring, W.J. (1988) EMBO J. 7, 2569–2578.
74 Whiteway, M. and Szostak, J.W. (1985) Cell 43, 483–492.
75 Prost, E., Deryckere, F., Roos, C. Haenlin, M., Pantesco, V. and Mohier, E. (1988) Genes Dev. 2, 891–900.
76 Mahaffey, J.W. and Kaufman, T.C. (1988) in Developmental Genetics of Higher Organisms. A Primer in Developmental Biology (Malacinski, G.M., ed.), pp. 329–360, Macmillan, New York.
77 Laughon, A., Carroll, S.B., Storfer, F.A., Riley, P.D. and Scott, M.P. (1985) Cold Spring Harbor Symp. Quant. Biol. 50, 253–262.
78 Frigerio, G., Burri, M., Bopp, D., Baumgartner and Noll, M. (1986) Cell 47, 735–746.
79 Felsenstein, J. (1988) Annu. Rev. Gen. 22, 521–565.
80 Baumgartner, S., Bopp, D., Burri, M. and Noll, M. (1987) Genes Dev. 1, 1247–1267.
81 Beeman, R.W. (1987) Nature 327, 247–249.
82 Kaufman, T.C. and Abbott, M.K. (1984) In Molecular Aspects of Early Development (Malacinski, G.M. and Klein, W.H., eds.), pp. 189–218, Plenum Press, New York.
83 Merrill, V.K.L., Turner, F.R. and Kaufman, T.C. (1987) Dev. Biol. 122, 379–395.
84 Mahaffey, J.W. and Kaufman, T.C. (1987) Genetics 117, 51–60.
85 Struhl, G. (1984) Nature 308, 454–457.
86 Tiong, S.Y.K., Whittle, J.R.S. and Gribbin, M.C. (1987) Development 101, 135–142.
87 Harvey, R.P., Tabin, C.J. and Melton, D.A. (1986) EMBO J. 5, 1237–1244.
88 Martin, G.R., Boncinelli, E., Duboule, D., Gruss, P., Jackson, I., Krumlauf, R., Lonai, P., McGinnis, W., Ruddle, F. and Wolgemuth, D. (1987) Nature (Lond.) 325, 21–22.
89 McGinnis, W., Hart, C.P., Gehring, W.J. and Ruddle, F.H. (1984c) Cell 38, 675–680.
90 Colberg-Poley, A.M., Voss, S.D., Chowdhury, K. and Gruss, P. (1985) Cell 43, 39–45.
91 Colberg-Poley, A.M., Voss, S.D., Chowdhury, K. and Gruss, P. (1985) Nature 314, 713–718.
92 Duboule, D., A. Baron, P. Mähl and B. Galliot (1986) EMBO J. 5, 1973–1980.
93 Gaunt, S.J., Miller, J.R., Powell, D.J. and Duboule, D. (1986) Nature 324, 662–664.
94 Wolgemuth, D.J., Engelmyer, E., Duggal, R.N., Gizang-Ginsberg, E., Mutter, G.L., Ponzetto, C., Viviano, C. and Zakeri, Z.F. (1986) EMBO J. 5, 1229–1235.
95 Rubin, M.R., Toth, L.E., Patel, M.D., D'Eustachio, P. and Chi Nguyen-Huu, M. (1986) Science 233, 663–667.
96 Baron, A., Featherstone, M.S., Hill, R.E., Hall, A., Galliot, B. and Duboule, D. (1987) EMBO J. 6, 2977–2986.
97 Odenwald, W.F., Taylor, C.F., Palmer-Hill, F.J., Friedrich, V., Jr., Tani, M. and Lazarini, R.A. (1987) Genes Dev. 1, 482–496.
98 Hart, C.P., Awgulewitsch, A., Fainsod, A., McGinnis, W. and Ruddle, F.H. (1985) Cell 43, 9–18.
99 Jackson, I.J., Schofield, P. and Hogan, B. (1985) Nature 317, 745–748.
100 Hauser, C.A., Joyner, A.L., Klein, R.D., Learned, T.K., Martin, G.R. and Tjian, R. (1985) Cell 43, 19–28.
101 Hart, C.P., Fainsod, A. and Ruddle, F.H. (1987) Genomics 1, 182–195.
102 Fibi, M., Zink, B., Kessel, M., Colberg-Poley, A.M., Labeit, S., Lehrach, H. and Gruss, P. (1988) Development 102, 349–359.
103 Graham, A., Papalopulu, N., Lorimer, J., McVey, J.H., Tuddenham, E.G.D. and Krumlauf, R. (1989) Genes Dev. 2, 1424–1428.
104 Papalopulu, N., Graham, A., Lorimer, J., Kenny, R., McVey, J. and Krumlauf, R. (1988) Structure, Expression and Evolutionary Relationships of Murine Homeobox Genes in the Hox2 Cluster. In Cell to Cell Signalling in Mammalian Development, NATO ASI Series, Springer Verlag, in press.
105 Lonai, P., Arman, E., Czosnek, H., Ruddle, F.H. and Blatt, C. (1987) DNA 6, 409–418.
106 Sharpe, P.T., Miller, J.R., Evans, E.P., Burtenshaw, M.D. and Gaunt, S.J. (1988) Development 102, 397–407.
107 Awgulewitsch, A., Utset, M.F., Hart, C.P., McGinnis, W. and Ruddle, F.H. (1986) Nature 320, 328–334.
108 Breier, G., Bućan, M., Francke, U., Colberg-Poley, A.M. and Gruss, P. (1986) EMBO J. 5, 2209–2215.
109 Joyner, A.L., Lebo, R.V., Kan, Y.W., Cox, D.R. and Martin, G.R. (1985) Nature 314, 173–175.
110 Levine, M., Rubin, G.M. and Tjian, R. (1984) Cell 38, 667–673.
111 Boncinelli, E., Simeone, A., La Volpe, A., Faiella, A., Fidanza, V., Acampora, D. and Scotto, L. (1985) Cold Spring Harb. Symp. Quant. Biol. 50, 301–306.
112 Rabin, M., Hart, C.P., Ferguson-Smith, A., McGinnis, W., Levine, M. and Ruddle, F. (1985) Nature 314, 175–178.
113 Rabin, M., Ferguson-Smith, A., Hart, C.P. and Ruddle, F.H. (1986) Proc. Natl. Acad. Sci. USA 83, 9104–9108.
114 Bućan, M., Yang-Feng, T., Colberg-Poley, A.M., Wolgemuth, D.J., Guenet, J.-L., Francke, U. and Lehrach, H. (1986) EMBO J. 5, 2899–2905.
115 Simeone, A., Mavilio, F., Bottero, L., Giampaolo, A., Russo, G., Faiella, A., Boncinelli, E. and Peschle, C. (1986) Nature 320, 763–765.
116 Boncinelli, E., Somma, R., Acampora, D., Pannese, M., D'Esposito, M. and Simeone, A. (1988) Organization of human homeobox genes. Hum. Reprod. 3, 880–886.
117 Cannizzaro, L.A., Croce, C.M., Griffin, C.A., Simeone, A., Boncinelli, e. and Huebner, K. (1987) Am. J. Hum. Genet. 41, 1–15.
118 Acampora, D., Pannese, M., D'Esposito, M., Simeone, A. and Boncinelli, E. (1987) Hum. Reprod. 2, 407–414.
119 Fienberg, A.A., Utset, M.F., Bogorad, L.D., Hart, C.P., Awgulewitsch, A., Ferguson-Smith, A., Fainsod, A., Rabin, M. and Ruddle, F.R. (1987) Curr. Topics Dev. Biol. 23, 233–256.
120 Laughon, A. and Scott, M.P. (1984) Nature 310, 25–31.
121 Pabo, C.O. and Sauer, R.T. (1984) Annu. Rev. Biochem. 53, 293–321.
122 Anderson, J.E., Ohlendorf, D.H., Takeda, Y. and Matthews, B.W. (1981) Nature 290, 754–758.
123 Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97–120.
123a Otting, G., Yan-qui, Q., Müller, M., Affolter, M., Gehring, W.J. and Wüthrich, K. (1988) EMBO J. 7, 4305–4309.
124 Casanova, J., Sanchez-Herrero, E. and Morata, G. (1988) EMBO J. 7, 1097–1105.
125 Rowe, A. and Akam, M. (1988) EMBO J. 7, 1107–1114.
126 McKay, D.B. and Steitz, T.A. (1981) Nature 290, 744–749.
127 Pabo, C.O. and Lewis, M. (1982) Nature 298, 443–451.
128 Anderson, J.E., Ptashne, M. and Harrison, S.C. (1985) Nature 316, 596–605.
129 Anderson, J.E., Ptashne, M. and Harrison, S.C. (1987) Nature 326, 846–852.
130 Schevitz, R.W., Otwinowski, Z., Joachinmiak, A., Lawson, C.L. and Sigler, P.B. (1985) Nature 317, 782–786.
131 Wharton, R.P. and Ptashne, M. (1985) Nature 316, 601–605.
132 Ebright, R.H., Cossart, P., Gicquel-Sanzey, B. and Beckwith, J. (1984) Nature 311, 232–235.
133 Wharton, R.P., Brown, E.L. and Ptashne, M. (1984) Cell 38, 361–369.

134 Kelley, R.L. and Yanofsky, C. (1985) Proc. Natl. Acad. Sci. USA 82, 483–487.

135 Ohlendorf, D.H., Anderson, W.F., Fisher, R.G., Takeda, Y. and Matthews, B.W. (1982) Nature 298, 718–723.

136 Ohlendorf, D.H., Anderson, W.F., Lewis, M., Pabo, C.O. and Matthews, B.W. (1983) J. Mol. Biol. 169, 757–769.

137 Matthews, B.W., Ohlendorf, D.H., Anderson, W.F., Fisher, R.G. and Takeda, Y. (1982) Cold Spring Harb. Symp. Quant. Biol. 47, 427–433.

138 Nelson, H.C.M., Hecht, M.H. and Sauer, R.T. (1982) Cold Spring Harb. Symp. Quant. Biol. 47, 441–449.

139 Hecht, M.H., Nelson, H.C.M. and Sauer, R.T. (1983) Proc. Natl. Acad. Sci. USA 80, 2676–2680.

140 Nelson, H.C.M. and Sauer, R.T. (1985) Cell 42, 549–558.

141 Lehming, N., Sartorius, J. Niemöller, M., Genenger, G., v. Wilcken-Bergmann, G. and Müller-Hill, B. (1987) EMBO J. 6, 3145–3153.

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

143 Fainsod, A., Bogarad, L.D., Ruusala, T., Lubin, M., Crothers, D.M. and Ruddle, F.H. (1986) Proc. Natl. Acad. Sci. USA 83, 9532–9536.

144 Odenwald, W.F., Garbern, J., Arnheiter, H., Tournier-Lasserve, E. and Lazzarini, R.A. (1988) The HOX1.3 homeo box gene encodes a sequence specific DNA binding phosphoprotein. In Cell to Cell Signals in Mammalian Development, NATO ASI Series, in press.

145 Hoey, T. and Levine, M. (1988) Nature 332, 858–861.

146 Desplan, C., Theis, J. and O'Farrell, P.H. (1988) Cell, in press.

147 Cho, K.W.Y., Goetz, J., Wright, C.V.E., Fritz, A., Hardwicke, J. and De Robertis, E.M. (1988) EMBO J. 7, in press.

148 Beachy, P. (1986) Ph.D. Thesis. Stanford University, Stanford, CA.

149 Beachy, P.A., Krasnow, M.A., Gavis, E.R. and Hogness, D.S. (1988) Cell 55, 1069–1081.

150 Müller, M., Affolter, M., Leupin, W., Otting, G., Wüthrich, K., and Gehring, W.J. (1988) EMBO J. 7, 4299–4304.

151 Laughon, A., Howell, W. and Scott, M.P. (1988) Development 104 (Suppl.), 75–83.

152 Johnson, A.D. and Herskowitz, I. (1985) Cell 42, 237–247.

152 a Krasnow, M.A., Saffman, E.E., Kornfeld, K. and Hogness, D.S. (1989) Cell, in press.

152 b Winslow, G.M., Hayashi, S., Krasnow, M.A., Hogness, D.S. and Scott, M.P. (1989) Cell, in press.

153 Bodner, M., Castrillo, J.-L., Theill, L.E., Deerinck, T., Ellisman, M. and Karin, M. (1988) Cell 55, 505–518.

154 Harvey, R.P., Robins, A.J. and Wells, J.R.E. (1982) Nucleic Acids Res. 10, 7851–7863.

155 Falkner, F.G. and Zachau, H.G. (1984) Nature 310, 71–74.

156 Parslow, T.G., Blair, D.L., Murphy, W.J. and Granner, D.K. (1984) Proc. Natl. Acad. Sci. USA 81, 2650–2654.

157 Fletcher, C., Heintz, N. and Roeder, R.G. (1987) Cell 51, 773–781.

158 Sturm, R., Baumruker, T., Franza, B.R., Jr. and Herr, W. (1987) Genes Dev. 1, 1147–1160.

159 Staudt, L.M., Singh, H., Sen, R., Wirth, T., Sharp, P.A. and Baltimore, D. (1986) Nature 323, 640–643.

160 Wirth, T., Staudt, L. and Baltimore, D. (1987) Nature 329, 174–178.

161 Staudt, L.M., Clerc, R.G., Singh, H., LeBowitz, J.H., Sharp, P.A. and Baltimore, D. (1988) Science 241, 577–580.

162 LeBowitz, J.H., Kobayashi, T., Staudt, L., Baltimore, D. and Sharp, P.A. (1988) Genes Dev. 2, 1227–1237.

163 Scheidereit, C., Heguy, A. and Roeder, R.G. (1987) Cell 51, 783–793.

163 O'Neill, E.A., Fletcher, C., Burrow, C.R., Heintz, N., Roeder, R.G. and Kelly, T.J. (1988) Science 241, 1210–1213.

165 Carroll, S.B., DiNardo, S., O'Farrell, P.H., White, R.A.H. and Scott, M.P. (1988) Genes Dev. 2, 350–360.

166 Tautz, D., Lehmann, R., Schnürch, Schuh, R., Seifert, E., Kienlin, A., Jones, K. and Jäckle, H. (1987) Nature 327, 383–389.

167 Schröder, C., Tautz, D., Seifert, E. and Jäckle, H. (1988) EMBO J. 7, 2881–2887.

167 a Driever, W. and Nüsslein-Volhard, C. (1989) Nature 337, 138–143.

168 Kilcherr, F., Baumgartner, S., Bopp, D., Frei, E. and Noll, M. (1986) Nature 321, 493–499.

169 Carroll, S.B. and Scott, M.P. (1986) Cell 45, 113–126.

170 Frasch, M. and Levine, M. (1987) Genes Dev. 1, 981–995.

171 Doe, C.Q., Hiromi, Y., Gehring, W.J. and Goodman, C.S. (1988) Science 239, 170–175.

172 Howard, K. and Ingham, P.W. (1986) Cell 44, 949–957.

173 Frasch, M., Hoey, T., Rushlow, C., Doyle, H. and Levine, M. (1987) EMBO J. 6, 749–759.

174 DiNardo, S. and O'Farrell, P.H. (1987) Genes. Dev. 1, 1212–1225.

175 Martinez-Arias, A. and White, R.A.H. (1988) Development 102, 325–338.

176 Ingham, P.W., Baker, N.E. and Martinez-Arias, A. (1988) Nature 331, 73–75.

177 Ingham, P.W. and Martinez-Arias, A. (1986) Nature 324, 592–597.

178 Hafen, E., Levine, M. and Gehring, W.J. (1984) Nature 307, 287–289.

179 Harding, K., Wedeen, C., McGinnis, W. and Levine, M. (1985) Science 233, 953–959.

180 Carroll, S.B., Laymon, R.A., McCutcheon, M.A., Riley, P.D. and Scott, M.P. (1986) Cell 47, 113–122.

181 Struhl, G. and White, R.A.H. (1985) Cell 43, 507–519.

182 Boulet, A.M. and Scott, M.P. (1988) Genes Dev. 2, 1600–1614.

183 Bodmer, R., Barbel, S., Sheperd, S., Jack, J.W., Jan, L.Y. and Jan, Y.N. (1987) Cell 51, 293–307.

184 Harvey, R.P. and Melton, D.A. (1988) Cell 53, 687–697.

185 Hogan, B.L.M., Holland, P.W.H. and Schofield, P.N. (1985) Trends Genet. 1, 67–74.

186 Thomas, K.R. and Capecchi, M.R. (1987) Cell 51, 503–512.

187 Herskowitz, I. and Y. Oshima (1981) in Molecular Biology of Yeast *Saccharomyces* (Strathern, N.J., Jones, E.A. and Broach, J.R., eds.), pp. 181, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

188 Strathern, J., Hicks, J. and Herskowitz, I. (1981) J. Mol. Biol. 147, 357–372.

189 Miller, A.M., McKay, V.L. and Nasmyth, K.A. (1985) Nature 314, 598–602.

190 Klar, A.J.S., Strathern, J.N., Broach, J.R. and Hicks, J.B. (1981) Nature 310, 25–30.

191 Nasmyth, K.A., Tatchell, K., Hall, B.D., Astell, C. and Smith, M. (1981) Nature 289, 244–250.

192 Shepherd, J.C., McGinnis, W., Carrasco, A.E., De Robertis, E.M. and Gehring, W.J. (1984) Nature 310, 70–71.

193 Sauer, R.T., Smith, D.L. and Johnson, A.D. (1988) Genes Dev. 2, 807–816.

194 Hall, M.N. and Johnson, A.D. (1987) Science 237, 1007–1012.

195 Bender, A. and Sprague, G.F., Jr. (1987) Cell 50, 681–691.

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

197 Goutte, C.A. and Johnson, A.D. (1988) Cell 52 875–882.

198 Siliciano, P.G. and Tatchell, K. (1984) Cell 37, 969–978.

199 Dolecki, G.J., Wannakrairoj, S., Lum, R., Wang, G., Riley, H.D., Carlos, R., Wang, A. and Humphreys, T. (1986) EMBO J. 5, 925–930.

200 Kessel, M., Schulze, F., Fibi, M. and Gruss, P. (1987) Proc. Natl. Acad. Sci. USA 84, 5306–5310.

201 Schughart, K., Utset, M.F., Awgulewitsch, A. and Ruddle, F.H. (1988) Proc. Natl. Acad. Sci. USA 85, 5582–5586.

202 Meijlink, F., DeLaaf, R., Verrijzer, P., Destrée, O., Kroezen, V., Hilkens, V., Hilkens, J. and Deschamps, J. (1987) Nucl. Acids. Res. 15, 6773–6786.
203 Falzon, M. and Chung, S.Y. (1988) Development 103, 601–610.
204 Falzon, M., Sanderson, N. and Chung, (1987) Gene 54, 23–32.
205 Kuroiwa, A., Kloter, U., Baumgartner, P. and Gehring, W.J. (1985) EMBO J. 4, 3757–3764.
206 Condie, B.G. and Harland, R.M. (1987) Development 101, 93–105.
207 Wright, C.V.E., Cho, K.W.Y., Fritz, A., Bürglin, T.R. and De Robertis, E.M. (1987) EMBO J. 6, 4083–4094.
208 Fritz, A. and De Robertis, E.M. (1988) Nucl. Acids Res. 16, 1453–1469.
209 Eiken, H.G., Njolstad, P.R., Molven, A. and Fjose, A. (1987) Biochem. Biophys. Res. Commun. 149, 1165–1171.
210 Featherstone, M.S., Baron, A., Gaunt, S.J., Mattei, M.-G. and Duboule, D. (1988) Proc. Natl. Acad. Sci. USA 85, 4760–4764.
211 Tournier-Lasserve, E., Odenwald, W.F., Garbern, J. and Lazzarini, R.A. (1988) The human cognate of the murine Hox 1.3 homeo box gene is almost identical to its murine counterpart. In Cell to Cell Signals in Mammalian Development, Springer Verlag, in press.
212 Breier, G., Dressler, G.R. and Gruss, P. (1988) EMBO J. 7, 1329–1336.
213 Mouellic, H.L., Condamine, H. and Brûlet, P. (1988) Genes Dev. 2, 125–135.
214 Rubin, M.R., King, W., Toth, L.E., Sawczuk, I.S. Levine, M.S., D'eustachio, P. and Nguyen-Huu, M.C. (1987) Mol. Cell. Biol. 7, 3836–3841.
215 Krumlauf, R. et al. (1988) EMBO J. 7, 3131.
216 Costa, M. and Kenyon, C. (1988) Cell 55, 747–756.
217 Miller, A.M. (1984) EMBO J. 3, 1061–1065.
218 Bürglin, T.R. (1988) Cell 53, 339–340.