Volume 10 Number 14 1982 Nucleic Acids Research

The ovalbumin gene family: complete sequence and structure of the Y gene

R.Heilig, R.Muraskowsky, C.Kloepfer and J.L.Mandel

Laboratoire de Génétique Moléculaire des Eucaryotes du CNRS, Unité 184 de Biologie Moléculaire et de Génie Génétique de l'INSERM, Faculté de Médecine de Strasbourg, Institut de Chimie Biologique, 11 rue Humann, 67085 Strasbourg Cédex, France

Received 2 June 1982; Accepted 23 June 1982

ABSTRACT.

The "ovalbumin Y" gene, one of three which constitute the ovalbumin gene family in chicken has been completely sequenced. The exact location of exons can be derived from the comparison with the ovalbumin gene sequence and from the map previously established by electron microscopy analysis. During evolution of the Y gene, selective pressure has operated to retain a sequence coding for an ovalbumin-like protein. The location of splice junctions, the length of protein coding exons and the reading phase are as in the ovalbumin gene. The overall homology between the Y and ovalbumin protein coding sequences is 72.6 % (resulting in a 58 % homology for the amino acid sequences). A significantly high number of base changes within coding sequences are present in clusters, which appear in several cases to be correlated with the occurence of direct repeats.

The 3' untranslated sequences of the Y and ovalbumin mRNAs have diverged much more, and the Y sequence contains a peculiar U(T) rich region. Corresponding introns of the ovalbumin and Y genes differ extensively both in sequence and in length. They share however characteristic biases in their base distribution.

INTRODUCTION.

The ovalbumin gene family in the chicken is composed of a cluster of 3 genes, the X, Y and ovalbumin genes, located within a 40 kb region (1, 2). These three genes are transcribed in chick oviduct under similar but non identical steroïd hormone control and code for proteins (3), although the X and Y proteins have not yet been identified *in vivo*. Comparison of the structures of the X, Y and ovalbumin genes has shown that they arose by duplication events from a common ancestor. Electron microscopy analysis of hybrids between the cloned genes and the corresponding RNAs has shown that all 3 genes are composed of 8 exons and the corresponding exons are of very similar size (2). Sequence homologies in exonic regions have been demonstrated by cross-hybridization experiments (1, 2) and by direct sequencing of 3 exons of the X gene (2). In order to analyse the modes of evolution and expression of such complex genes, we have undertaken a detailed

0305-1048/82/1014-4363$2.00/0

structural study of the X and Y genes. We report here the complete sequence of the Y gene and its comparison with the ovalbumin gene sequence.

MATERIALS AND METHODS.

Sequencing strategy.

Almost all of the Y gene sequence presented in Fig. 2 was obtained from 5' end-labelled, strand-separated fragments, since sequences performed in this way are of higher quality than those obtained from double stranded fragments labelled at one end (they are sharper and less subject to artefacts in the pyrimidine tracks). In the strategy used, we did not try to first establish a detailed restriction map. For each subclone we looked for restriction nucleases which would produce a pattern of well spaced fragments ranging from about 100 to 800 bp. Such a mixture of fragments was end-labelled, submitted to strand separation, and the isolated single stranded fragments were then sequenced. Using two sets of fragments produced with different enzymes, we obtained overlapping sequences covering 80 to 95 % of the subclone. This generated a restriction map (although incomplete) which guided the further experiments required to fill in the gaps.

Isolation of single stranded DNA fragments.

This was performed following the Maxam and Gilbert protocol (4), with some modifications. Restriction enzyme digests (about 30 pmols of ends) were treated with bacterial alkaline phosphatase (Worthington), phenol extracted, ethanol precipitated and 5' end-labelled with T4 polynucleotide kinase (Boehringer) according to Maxam and Gilbert (4). The reaction was stopped with SDS (0.1 %) and EDTA (10 mM), ethanol precipitated twice, and the dried DNA pellet was well dissolved in 40 µl of 2 mM EDTA (containing xylene cyanol). 60 µl of 99 % DMSO was then added. The sample was heated for 5' at 90°C, quickly cooled and subjected to electrophoresis (18 mA at 4°C) on a polyacrylamide gel (8 to 12 %, depending on fragment sizes, with an acrylamide/bisacrylamide ratio of 60/1). Electrophoresis and gel buffer was tris-borate 50 mM, EDTA 1 mM, pH 8.3 (4). Bands were localised by autoradiography, cut out and crushed as described by Maxam and Gilbert(4). Acrylamide was pelleted by centrifugation, the supernatant was filtered through siliconized glass wool and precipitated with ethanol, using tRNA (20 µg) as carrier. Because of the presence of this tRNA, we did not add calf thymus DNA in the sequence reactions as proposed by Maxam and Gilbert (4).

RESULTS

DNA sequence of the Y gene.

The Y gene has been entirely sequenced by the Maxam and Gilbert technique (4) as outlined in Materials and Methods. For most of this work we used subclones (described in Fig.1b) derived from the cosmid clone pAR2 (1) However the leader sequence and the 5' half of intron A were sequenced from a different clone (clone λXEco10, see Fig. 1c). Although both the cosmid clone and clone XEco10 were obtained using DNA from chickens of the same flock, comparison of sequences in intron A derived from the two clones shows that they differ by about 1 % (see Fig. 2). Within the various subclones, all sites used for 5' end-labelling were overlapped in a different sequence. However in 2 cases we did not sequence across sites corresponding to the limits of HindIII subclones (see Fig. 1b). These two HindIII sites are found within introns. Since they are 6 bp sites, their average frequency is 1 in 3000 bp (taking into account base composition) and the chance of overlooking a small fragment (< 100 bp) is low. The excellent fit between intron lengths derived from the sequence and the previous electron microscopy measurements rules out the possibility that longer fragments have been

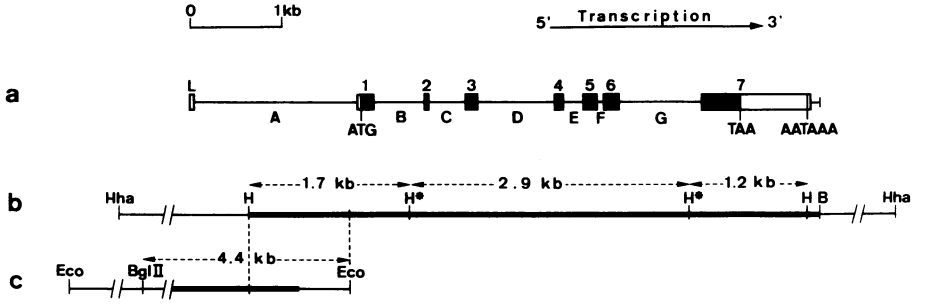
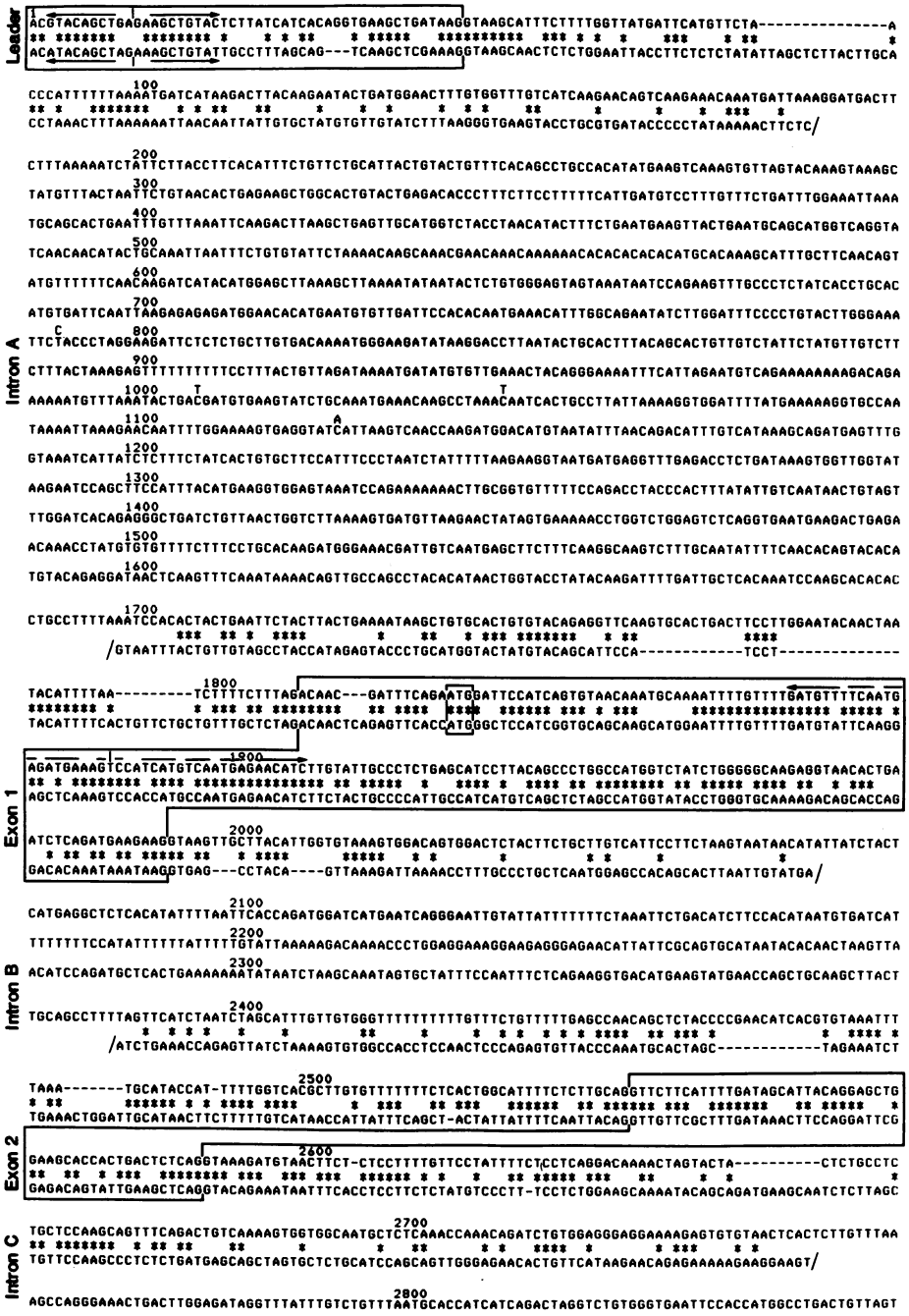
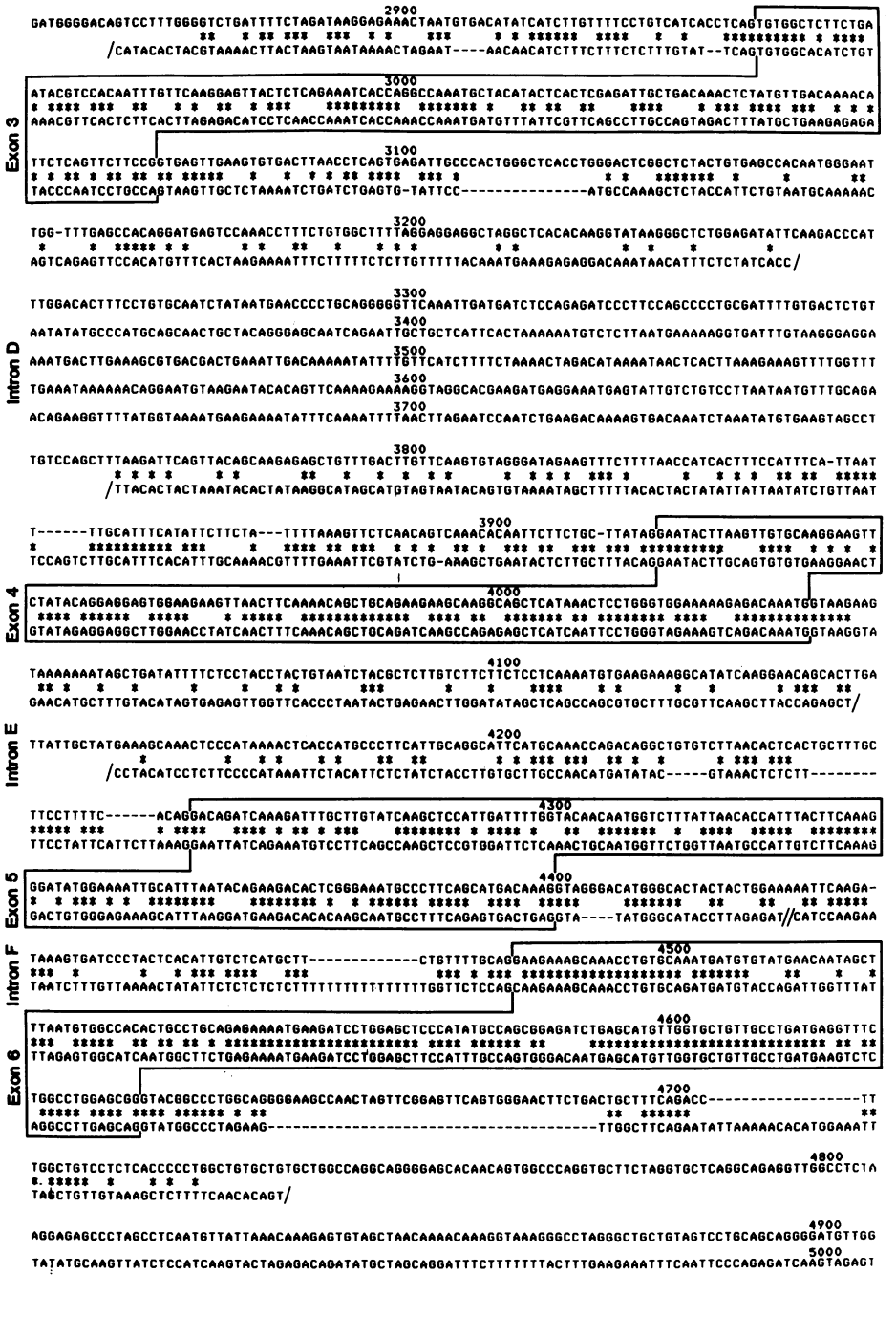
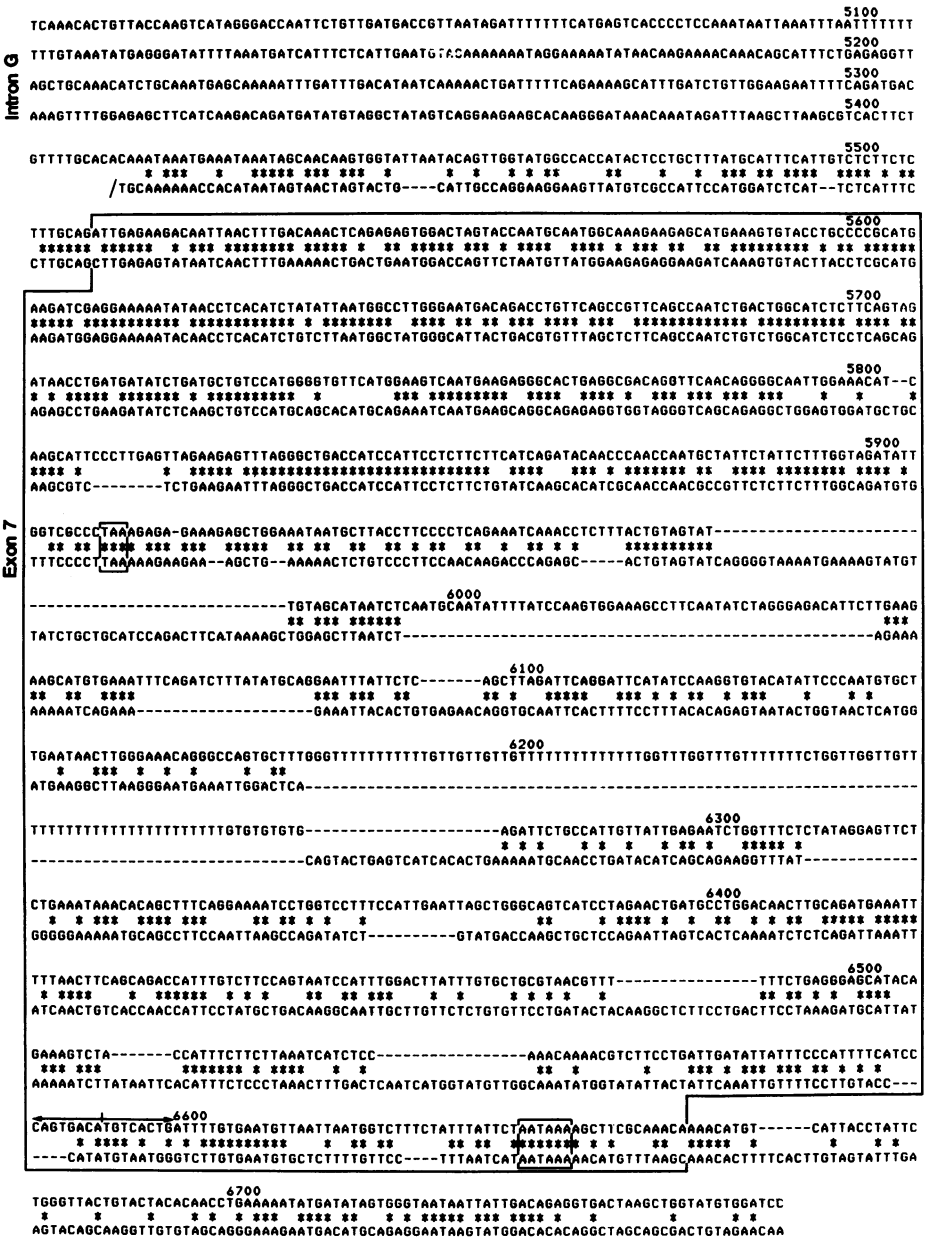


Figure 1 : Map of the Y gene. a) Position of exons (L and 1 to 7). White boxes : untranslated regions, black boxes : protein-coding regions. ATG and TAA correspond to the initiation and termination codons and AATAAA to the "polyadenylation signal" (see text). Introns are labelled A to G. The limits of the map correspond to those of the sequence presented in Fig. 2. b) Map of subclones derived from the large Hha fragment of the cosmid clone pAR2 (1). H and B stand for HindIII and BamH1 respectively. The heavy line corresponds to the sequenced region of these clones. We did not sequence across the starred restriction sites. c) Location of the EcoRI-BglII subclone derived from the λXEco10 clone (see Ref. 1). The heavy line represents the sequenced region. Sequences upstream from the HindIII site and from the leader exon have been published previously (5).


Leader
Intron A
Exon 1
Intron B
Exon 2
Intron C


Exon 3
Intron D
Exon 4
Intron E
Exon 5
Intron F
Exon 6


Intron G
Exon 7
5100
5200
5300
5400
5500
5600
5700
5800
5900
6000
6100
6200
6300
6400
6500
6600
6700

missed (Table I). The 6759 bp sequence presented in Fig. 2 starts at the cap site for Y gene transcripts which we have localised previously (5) and ends about 120 bp beyond the polyadenylation site (see below).

Alignment with the ovalbumin gene sequence.

a) Conservation of sequences involved in post-transcriptional processing.

Sequences showing a high degree of homology to all of the ovalbumin exons can be found in the Y gene (Fig. 2). The alignment is unambiguous except for a major part of the 3' untranslated sequence in exon 7 (corresponding to positions 5979 to 6585 in the Y gene sequence). The sequences defining splice sites in the ovalbumin gene are well conserved in the Y gene and conform in all cases to consensus sequences for such sites (see Ref. 7 and 8). The region surrounding the polyadenylation site in the ovalbumin gene can be aligned with a homologous sequence in the Y gene which includes the AATAAA sequence (at position 6636), commonly found at the 3' end of protein-coding genes (9, 10). Assuming that these homologous sequences are functional in the maturation of the Y gene transcripts, it is possible to define the lengths of the exons and introns. These lengths fit very well with those previously determined by electron microscopy analysis of hybrids between Y mRNA and the cloned DNA (Table I).

b) Protein-coding sequences.

A putative Y mRNA sequence can be deduced and analysed for the location of an initiation codon and for the distribution of stop codons. The first AUG in this sequence is present at a position in exon 1 homologous to the initiation codon for ovalbumin and defines a reading frame which is open up to a UAA codon, 1164 nucleotides downstream, which is also homologous to the ovalbumin termination codon (position 5911 in exon 7). Throughout this interval, the two other phases contain many stop codons : 20 and 32 for the 2nd and 3rd phases respectively (Table I). Following the first UAA in phase 1, many nonsense codons can be found in all three phases

Figure 2 : The Y gene sequence and its alignment with homologous regions in ovalbumin gene. The Y gene sequence is presented above the corresponding ovalbumin gene sequence (6). The numbering corresponds to the Y gene sequence. From position 1 to 618, the sequence is derived from λXEcoI0 clone and past this position, from the pAR2 clone. In the region 618 to 1158, which has been sequenced on both clones (see Fig. 1), we have indicated above the line the bases which differ in XEcoI0. For non-coding regions, the alignment proposed is based on a minimum number of deletions. In intron regions which show no significant homology only the Y gene sequence is shown. The exons, the initiation and termination codons and the AATAAA polyadenylation signal are boxed. Palindromic sequences in the leader, in exon 1, and upstream the polyadenylation site are indicated by arrows.

TABLE I : Length of structural elements in Y and ovalbumin genes and distribution of nonsense codons in Y gene exons.

Structural elements	Ov. gene (bp)	Y gene (bp)		Non-sens codons (Ygene) phase1	phase2	phase3
leader exon (untranslated)	47	50		1	3	0
intron A	1589	1757	(1619±145)			
exon 1 -untranslated region	17	14	(199±22)	0	0	0
-coding region	168	168		0	4	5
intron B	251	545	(453±98)			
exon 2	51	51	(54±15)	0	0	3
intron C	581	355	(357±56)			
exon 3	129	129	(142±16)	0	0	3
intron D	400	847	(808±83)			
exon 4	118	118	(128±17)	0	2	1
intron E	958	221	(235±35)			
exon 5	143	143	(148±16)	0	1	3
intron F	331	81	(50±15)			
exon 6	156	156	(165±18)	0	4	4
intron G	1582	876	(886±91)			
exon7 -coding region	393	399	(1146±76)	0	9	13
-untranslated region	650	744		10	15	10

For the Y gene, the numbers in parenthesis correspond to the length measurements obtained by electron microscopy analysis of mRNA-DNA hybrids (2). Phase 1 corresponds to the ovalbumin reading frame.

up to the AAUAAA sequence, as expected for an untranslated region. The protein-coding sequence thus defined in Y mRNA is almost identical in length to that in ovalbumin mRNA (1158 nucleotides), and is read in the same phase, except for a small region located 120 to 100 nucleotides upstream from the termination codon, where two insertion/deletion events appear to have taken place (positions 5801 to 5818). This is in excellent agreement with the results of *in vitro* translation, which showed that purified Y mRNA codes for a protein which has an electrophoretic mobility very close to that of ovalbumin in SDS polyacrylamide gel (3). These findings reinforce our conclusion that the exon limits in the Y gene are those determined on the basis of sequence homologies with the ovalbumin gene.

c) Non protein-coding sequences.

In the short (64 bp) 5' untranslated region, two small lengths differences (of 3 bp each) are found between the Y and ovalbumin gene sequences (in the leader and in exon 1, see Fig. 2). In the 3' untranslated region, alignment of the Y and ovalbumin sequences is not unambiguous. Homologous sequences are found somewhat scattered between dissimilar ones, and include a short stretch beyond the TAA codon and another one in the polyadenylation region. A peculiar TG-rich sequence, where stretches of T alternate with repeats of GTT, GGTTT, GGTT and GT, is found in the Y gene (positions 6171 to 6273) and has no counterpart in the corresponding ovalbumin region. Numerous insertions and deletions of a wide size range have to be introduced in order to optimize homology (17 length changes for the alignments proposed in Fig. 2). Related sequences are also found within the 100 bp which follow the polyadenylation site.

We did not find significant homology between Y and ovalbumin gene introns except for rather short segments adjacent to the splice junctions (Fig. 2). The length of the sequences that can be aligned in these regions ranges from 17 bp to 86 bp excluding the insertions/deletions which are needed to optimize the homology. Comparison of intron A sequences derived from two clones (originating from different chickens of the same flock) suggests that polymorphic differences are quite frequent (4 differences in 540 bp), as found previously for the ovalbumin gene (6, 11).

The Y protein sequence

From the Y mRNA coding sequence, it is possible to derive the sequence of a protein, 388 amino acid-long, which can be compared to the 386 amino acid-long ovalbumin sequence (Fig. 3). The putative Y protein is 58 % homologous to ovalbumin, which is not unexpected in view of the homology between the two nucleotide sequences, the identity of the reading phases (Fig. 2) and the better conservation, within codons, of replacement sites versus silent sites (see below, Table IV). The two additional amino acids in the Y protein are located near the carboxy-terminus (positions 352 and 357 in Fig. 3). Since little is known about functional or structural domains within the ovalbumin protein, the significance of the distribution of conserved and modified sequences cannot be assessed. Some long amino acid tracks are conserved (especially in the C terminal half, corresponding to exons 6 and 7) which are longer than expected from a random distribution of homologous amino acids (2).Several sites of post-translational modification are known in ovalbumin (12). The glycosylation site Asn(CHO)-Leu-Thr at

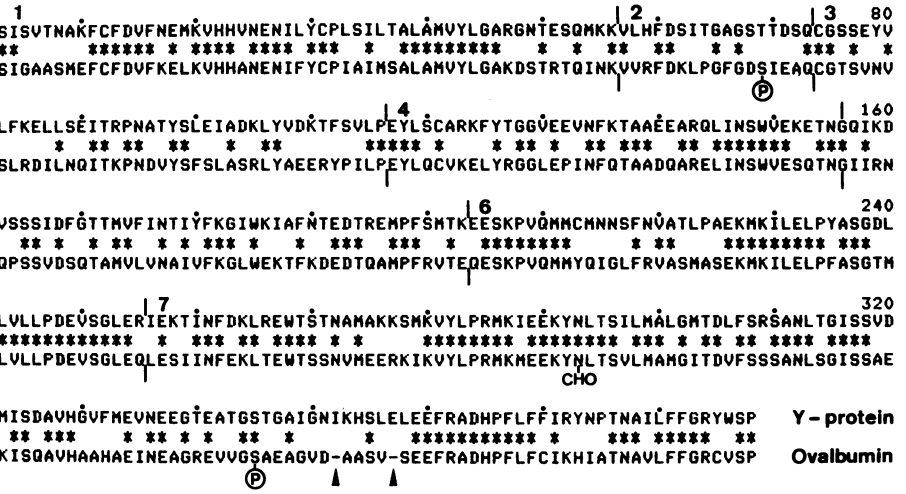


Figure 3 : Alignment of Y and ovalbumin protein sequences. The location of phosphorylation sites (P) and glycosylation site (CHO) are indicated for the ovalbumin sequence. Vertical bars delimitate sequences coded by exons 1 to 7. The two length differences between the two sequences are marked by arrowheads.

position 293 to 295 in ovalbumin is conserved in the Y protein, as well as one of the serine (aa 345) which can be phosphorylated in ovalbumin. The other phosphorylable serine (aa 69) is replaced by threonine in the Y sequence. Lingappa et al. (13) have provided some evidence that the ovalbumin region between aa 235 and 254 might play the role of non processed signal peptide involved in the secretion of the protein. This region is highly conserved in both Y and X protein sequences (Fig 3 and Ref. 2).

Non random base distribution in the Y and ovalbumin genes.

We have studied various parameters of base distribution in the ovalbumin and Y genes : base composition, dinucleotide frequency and occurence of polypurine or polypyrimidine stretches. Both genes have a base composition close to that found for the total chicken genome (14). Introns are more AT rich than protein-coding exons (Table II). Dinucleotide frequency has been analysed previously for the ovalbumin gene and it was reported that the only departure from the frequencies expected from base composition occured for the CG dinucleotide (6). A very low frequency of the CG pair is also found in the Y gene, in both exons and introns. However significant deviations from a random distribution are also found for some other

TABLE II : Parameters of base distribution in Y and Ovalbumin genes.

		Y GENE				OVALBUMIN GENE			
		Introns		Protein-coding region		Introns		Protein-coding region	
Base composition (%)	A	31,7		31,0		33,1		29,3	
	T	31,2		26,0		30,7		25,6	
	C	18,3		20,4		18,2		21,7	
	G	18,8		22,6		18,0		23,4	
Dinucleotide		Obs.	Exp.	Obs	Exp.	Obs.	Exp.	Obs.	Exp
composition	AT	362	(463)442	96	88	479	554	85	81
	TA	301	(463)442	49	88	465	554	41	81
	CG	18	(161)*	11	(54)*	29	(186)*	13	(59)*
	GC	164	(161)154	55	51	223	178	69	55
	AA	542	(470)448	116	105	670	599	101	93
	TT	526	(456)435	82	74	556	514	75	71
	CC	149	(157)189	49	59	168	224	52	68
	GG	190	(165)198	63	71	177	219	55	77
	AC	268	(272)259	61	69	318	329	52	69
	GT	244	(275)262	50	64	293	302	57	65
	AC+GT	512	(547)521	111	133	611	631	109	134
	CA	358	(272)328	101	90	419	406	108	92
	TG	361	(275)330	100	82	400	374	101	84
	CA+TG	719	(547)658	201	172	819	780	209	176
	AG	310	(279)334	89	99	417	402	102	96
	CT	332	(267)322	75	77	420	377	79	79
	AG+CT	642	(546)656	164	176	837	779	181	175
	GA	281	(279)266	95	77	330	325	90	74
	TC	276	(267)255	72	58	327	306	78	60
	GA+TC	557	(546)521	167	135	657	631	168	134

Base composition (in %) and frequency of dinucleotides have been calculated separately for introns and protein coding sequences (3' untranslated sequences are omitted since they are too short to yield significant results). The observed number of dinucleotides (obs) is compared to the number expected (exp) from base composition and corrected for the actual CpG frequency (see text). The uncorrected values are given (in parenthesis) for Y gene introns only and in all cases for the CG pairs (starred values). The statistical significance of the differences between observed and "corrected" expected values was calculated (except for CG pairs where the uncorrected value was taken into account). Boxed values indicate significance at $p<.02$.

dinucleotides (Table II). The complementary CA and TG pairs appear more frequently in coding and non coding sequences than expected from base composition alone, in both the ovalbumin and the Y genes (Table II, compare to AC and GT frequencies). Such a bias has been found in other vertebrate genomes and is correlated to the CG deficiency (15). It has been proposed that this phenomenon is due to a high mutation frequency of methylated cytosine towards thymine. In genomes where cytosine methylation is frequent at CG pairs (as in the chicken) this increase in mutation rate would lead both to a deficiency in the CG pair and to an increase in the TG and CA pairs derived from the C→T transition (15).

However, this interpretation is valid only if the increase in frequency of CA and TG pairs is mathematically independent of the decrease in CG pairs. This is not the case since the number of C (or G) residues in a sequence can be written as the sum of the number of dinucleotide pairs starting with C (or ending with G) : $N_C=N_{CA}+N_{CC}+N_{CG}+N_{CT}$ (and $N_G=N_{AG}+N_{CG}+N_{GG}+N_{TG}$). Thus the reduction in CG pairs will result in an increase in the expected number of CA, CC and CT pairs (and a similar effect will occur for AG, GG and TG pairs). When this is taken into account, we find that the observed frequency of CA and TG pairs is still higher than the expected values corrected for the CG bias (Table II). This would support the increased mutation hypothesis, but with an effect much smaller than previously reported (see for instance the Y gene introns, which have 719 CA+TG pairs compared to 547 expected from base composition alone, and to 658 after correction for the low CG content).

Other significant biases are found for A and T containing pairs. There are far fewer AT and TA pairs than AA or TT pairs in Y or ovalbumin gene introns. In protein coding exons however only the TA frequency deviates from expectation, albeit strikingly. Avoidance of TAA and TAG nonsense codons in the reading phase can explain only part of this bias. CC and GG appear also somewhat reduced and other deviations from expected values do not appear as common features in the two genes.

Examination of intron sequences in both the Y and ovalbumin genes reveals conspicuous runs of purines or pyrimidines. As shown in Table III there are more nucleotides in A+G stretches or C+T stretches for both size classes analysed, than expected. Frequency of polyA or polyT homopolymeric runs is increased in the Y gene only. Finally it can be noted that the YmRNA contains several interesting palindromes : in the leader (this palindrome is also present in the ovalbumin mRNA, but not in the XmRNA), in

TABLE III: Purine and Pyrimidine stretches in Y and ovalbumin gene introns.

Length of stretches (in bases)		Number of nucleotides in stretches					
		A+G stretches	C+T stretches	A_n	T_n	C_n	G_n
Y gene	6 to 10	417 (252)	372 (228)	60(21)	84(19)	0(1)	0(1)
	> 11	72 (15)	92 (12)				
Ovalbumin gene	6 to 10	479 (327)	323 (264)	28(32)	14(21)	6(1)	0(1)
	> 11	113 (21)	138 (14)		16(0,1)		

We counted the number of polypurine, polypyrimidine or homopolymeric tracts of a given length using the SEQFIT program of Staden (16). We then derived the total number of nucleotides present in tracts of two size classes : 6 to 10 nucleotides long and 11 nucleotides and longer. These observed values are compared to those expected from the base composition of the sequences (in parenthesis).

exon 1, and in exon 7 immediately upstream from the AAUAAA polyadenylation signal (see Fig. 2).

DISCUSSION.

a) The Y gene structure.

The location of exons in the Y gene, as presented here on the basis of sequence homology with the ovalbumin gene, is in excellent agreement with the previous electron microscopy results and allows the definition of a protein coding sequence of the length expected from *in vitro* translation of Y mRNA (3). However, since we did not compare the Y mRNA (or cDNA) sequence directly to the genomic sequences, we cannot *a priori* eliminate the possibility that at least some of the intron-exon junctions in the Y gene do not correspond to the ovalbumin gene junctions. From the study of a β^+ thalassemic globin gene it is known, in fact, that a single point mutation can create a new functional splice sequence within an intron, even though the sequence of the normal junction (and of the rest of the gene) has not been modified (17, 18, 19). If other splice junctions exist in the Y gene, they should be placed close to the one we have proposed (in order to accomodate the electron microscopy data) and should respect the reading frame, (since stop codons are found in the two other phases (Table I). Examination

of the Y gene did not reveal other possible junctions conforming to the proposed consensus sequences (8), that would satisfy these additional conditions. We feel therefore that the existence of an alternate structure for the Y gene is very improbable.

The location of the Y gene polyadenylation signal based on homology with the ovalbumin gene sequence is certainly accurate since it gives an excellent fit with the previous length determination for exon 7 (see Table I) and since no other polyadenylation signals (AATAAA, ATTAAA or AATATA, see Ref. 20, 21) are found in a 250 bp region around the proposed site. However the existence of minor polyadenylation sites cannot be eliminated : in fact a longer transcript has been detected which contains sequences beyond the proposed 3'end of the Y gene (3).

Comparison of the Y and ovalbumin gene structures shows that signals necessary for transcription initiation (TATA box, cap site, see Ref. 5) and post-transcriptional processing (splice sites and polyadenylation site, see above) have been conserved together with exon sequences. On the contrary, introns are generally very dissimilar in length as reported previously (Ref. 2, see Table I) and show very little sequence homology.

b) Selective pressure on Y protein-coding sequence.

The comparison of the ovalbumin and Y gene sequences confirms our previous observations (2) based on electron microscopy analysis of X, Y and ovalbumin gene structures and on examination of the sequence of 3 exons and neighbouring sequences in the X gene, that selective pressure was exerted mostly to preserve a protein-coding sequence. Little or no homology is found in introns or in most of the 3' untranslated sequence and even in the Y and ovalbumin non protein-coding regions which can be aligned, deletions or insertions appear to have occured at high frequency. In the 70 bp 5' untranslated sequence, two such events can be detected. In the 3' untranslated sequence, deletions or insertions occur as close as 5 bp downstream from the termination codon and we could align, with a 53 % homology, only 281 bp of the Y and ovalbumin 3' untranslated regions which are respectively 744 bp and 650 bp long. 17 deletions/insertions have to be introduced in either sequence to obtain such a result and we cannot therefore eliminate the possibility that some of the homology is due to chance.

In contrast, the conservation of splice junctions and the scarcity of insertion/deletion events within the coding sequences were essential to keep the reading frame. In the 1164 nucleotides of coding sequence the overall homology is 72,6 %, and only two closely located deletion/insertion

events appear to have taken place between the Y and ovalbumin genes, 100 bp upstream from the termination codon : these two events have compensated each other to restore the proper reading frame but resulted in an insertion of two amino acids in the Y sequence compared to the ovalbumin one (see Fig. 3, position 352 and 357). No such length differences are found between the protein-coding sequences of the X and Y genes. Examination of partial amino acid sequences from goose, pheasant, turkey and duck ovalbumins (22) shows that in this carboxy terminal region, they ressemble chicken ovalbumin and not the X and Y proteins. There is thus no evidence that genes orthologous to the X and Y chicken genes represent the major ovalbumin coding genes in these other avian species.

It is apparent that selective pressure has operated not only to keep an open reading frame, but also to preserve some common polypeptide structure between the ovalbumin and Y proteins. This is best indicated by a comparison of the mutation frequency at silent sites (where a base change does not modify the amino acid sequence) and at replacement sites (Table IV). Using the calculation proposed by Perler et al (23), we found, for the entire protein-coding sequences 29 % replacement changes versus 65 % silent changes (these results are corrected to account for multiple mutations at the same site).

Short regions of high homology are found in exon 1 (88 % homology over 67 bp) and in exon 4 (80 % over 69 bp) [see Fig. 2] which might be due to stronger selective pressure at the protein level since the two exons are characterized by a lower proportion of replacement changes, but not of silent changes (Table IV). More significantly, exons 6 and 7, which code for the entire carboxy-terminal half of the protein are also better conserved, as judged from the low level of changes at replacement sites (22 %) and from the occurence of long conserved amino acid tracks (> 10 aa, see above). In this case however, the proportion of silent changes appears to be lower than for the rest of the coding region (57 % compared to 73 %). Since silent mutations are thought to occur at a fixed rate, this latter observation might indicate that a correction event occured, relatively early in the history of Y and ovalbumin gene duplication. Decisive evidence for gene conversion events between the Y and X genes has been obtained by comparison of the two gene sequences (to be published).

The duplication of the Y and ovalbumin genes can be approximately dated to 50 to 80 million years ago, if one assumes that base substitutions at silent sites occur in avian at a rate similar to the value estimated for

TABLE IV : Comparison of Y and Ovalbumin Exons.

EXON	HOMOLOGY #	HOMOLOGY %	NUMBER OF GAPS (c)	REPLACEMENT SITES (% changes)	SILENT SITES (% changes)	
5'untranslated	42/61(a)	68	2(6)			
exon 1(coding region)	123/168	73	0	25,2	87,1	
exon 2	34/51	67	0	47,7	(46,1)	
exon 3	79/129	61	0	43,4	92,3	73,6
exon 4	86/118	73	0	24,4	72,6	
exon 5	96/143	67	0	42,1	69,5	
exon 6	122/156	78	0	19,8	50,2	57,6
exon 7 (coding region)	292/391(a)	75	2(10)	22,3	61,1	
3'untranslated (b)	281/528(a)	53	17(338)			

(a)For nucleotides which can be aligned
(b)The total length of the 3' untranslated region is 744bp in the Y gene and 650bp in the Ov gene, assuming that the site of polyA addition is at the same position with respect to the AATAAA sequence.
(c)The total length of the gaps is given in parenthesis.

Changes at replacement and silent sites have been calculated as in Perler et al. (23). Incidence of stochastic variation on the values obtained for individual exons is larger in the case of silent changes, since silent sites are less numerous than replacement sites by a factor of 2,5. For this reason, we provide in addition the values calculated for larger regions (a non linear correction for multiple changes is used in the calculation, and the value corresponding to a group of exons is thus not equivalent to the weighted average of the values corresponding to individual exons). Because of the small size of exon 2, the value for silent changes is not significant and is placed in parenthesis.

mammals ; 0.8 to 1.3 % changes per million years (see Ref. 24). In fact comparison with the limited amino acid sequence available on duck ovalbumin (22) suggests that there might be more differences between the Y protein and ovalbumin of either chicken or duck, than between the two latter proteins, which diverged for a period of about 80 million years.

c) Clustering of mutations in protein-coding exons

We have analysed the distribution of base differences between the Y and ovalbumin protein-coding exons, and found that clusters of at least four adjacent base changes are more frequent than expected from the 72,6 %

average homology (Fig.4a). This might be due to an absence of selective constraint in some regions where mutations would be fixed at a rate close to that for silent sites (see for instance exon 3). However some of the longer clusters might have arisen through other mechanisms : a striking example is given by the cluster of 8 consecutive base changes in an otherwise very conserved region in exon 7. At this place we found a 8/9 nucleotide adjacent direct repeat in the Y sequence and a related but different 7/7 nucleotide repeat in the ovalbumin gene (Fig. 4b). Other, less striking examples are the 6 bp long cluster in exon 1, where 4 bp repeats are found both in Y and in ovalbumin genes and the 5 bp long cluster in exon 1, where a 5/6 repeat is found in the ovalbumin gene sequence (Fig. 4 c and d). It has been previously proposed that deletions or insertions arise more frequently in regions where short duplications are found, due to mispairing during replication (25, 26). The sequences shown in Fig. 4b-d suggest that direct repeats might also play a role in the generation of block mutations.

d) Sequence features in non-coding regions.

Base composition of protein-coding regions is subjected to constraints related to the amino acid sequence. Intron sequences, although independent of such constraints, nevertheless have a non random base arrangement (see

a) Length of clusters (bp)	Observed	Expected I	Expected II
4	9	3,37	3,89
5	2	0,92	1,20
6	1	0,25	0,38
7	0	0,07	0,12
8	1	0,02	0,04
Number of base pairs in clusters ≥ 4	60	20,2	25

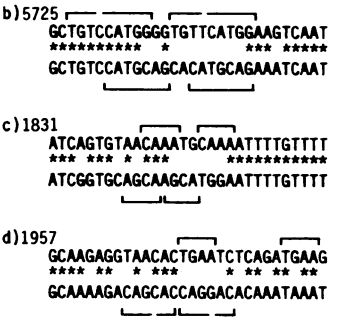


Figure 4 : Clustering of mismatches in protein-coding regions. a) The observed frequency of mismatch clusters of a defined length was compared to the expected values. The latter were calculated in two different ways : first on the basis of the 72,6 % homology found between the two coding regions and second by assuming a heterogeneous partition of the homology (67,5 % homology over 60 % of the length, 80,5 % homology for the remaining part). This assumption is supported by an analysis of homology over contiguous 50 bp long tracks (not shown). b to d) Sequences surrounding long clusters of mismatches. The direct repeats (see text) are indicated by square brackets. The Y gene sequence is presented above the corresponding ovalbumin gene sequence. Numbering is as in Fig. 2.

Table III and IV). The deficiency in CpG pairs is common to vertebrate genomes (introns and exons alike) and is related to the occurrence of cytosine methylation (15). Excess of TG and CA pairs is correlated to the CG deficiency, but only part of this excess might be due to increased mutation from CG to CA and TG (see Table III and above). The deficiency in AT and TA, and the excess in TT and AA pairs might be a characteristic of chicken and of other vertebrate genomes (27). It remains to be seen whether the increased occurrence of polypurine and polypyrimidine stretches found in introns is common to other genes or is specific for the ovalbumin gene family region and whether they have a biological significance. An asymmetric distribution of bases in the human γ globin genes and in other genes has recently been found and analysed by Smithies et al. (26), although the biological significance of such "chromosomal domains" is presently unknown.

A striking example of non random base distribution is given by the peculiar 103 bp long TG rich sequence found in the 3' untranslated region in exon 7. This sequence consists of stretches of Ts separated by repeats of small T+G motives and is not present in the ovalbumin gene. This sequence does not appear to be highly repeated in the genome (2). Another unusual 300 bp long sequence, which is TC rich, has been found upstream from the promoter region of the Y gene (5). Thus almost all of the Y gene (except its 3'end) is found between these two highly asymmetrical and relatively simple sequences. It has been proposed that simple sequences could favour gene rearrangements, such as the gene conversion which took place in the human γ globin locus (28, 29). However, there is no evidence for such a role in the Y gene, since similar levels of homology with the ovalbumin sequence are found, in the 3' untranslated region, on both sides of the TG rich sequence. Although such sequences might arise as "selfish DNA", their peculiar base composition could also be important in recognition by chromatin proteins. It is likely for instance that chromatin structure would be altered in these regions since it has been shown that nucleosomes do not form (in vitro) on dA-dT stretches (30, 31). AT rich spacers have been found as regular features of various genomes and it has been proposed that they play a role in the structuring of chromatin domains (32).

CONCLUSION.

The conservation of sequences important for the synthesis of a Y gene transcript, its maturation into a functional messenger, together with the apparent pressure to retain certain features in the protein sequence indi-

cate that the Y gene, although expressed at a low level, confers a distinct selective advantage to the chicken. No pressure appears to be exerted on the exact sequence of non coding regions, especially in introns, as has been observed for other gene families (24, 25). Even intron length evolved rapidly, contrary to the case of the globin gene family where these lengths appear to be conserved over longer evolutionary times (24). Non-random sequence arrangements are however found in common in introns of both Y and ovalbumin genes. Some of these biases might be general characteristics of chicken (or vertebrate) genomes. The most evident feature (concerning the CG pairs, but also the CA and TG pairs) is clearly related to DNA methylation. The biological significance of the other sequence features which have been found remains to be established.

ACKNOWLEDGEMENTS.

We wish to thank Prof. Chambon for his constant encouragement and support, Mrs. C. Tolstoshev and Mr. R. Fritz for designing the computer programs used, and J.M. Garnier for preparation of plasmids, B. Boulay and E. Badzinski for the preparation of the manuscript and Drs. Avner, Corden, Williams for editorial help. This work was supported by grants from the CNRS (ATP 006520/50 and ATP 4160), from the Association pour le Développement de la Recherche sur le Cancer, from the Fondation pour la Recherche Médicale Française and from the Fondation Simone et Cino del Duca.

REFERENCES.

1. Royal, A., Garapin, A., Cami, B., Perrin, F., Mandel, J.L., LeMeur, M., Brégégère, F., Gannon, F., LePennec, J.P., Chambon, P. and Kourilsky, P., (1979) Nature 279, 125-132.
2. Heilig, R., Perrin, F., Gannon, F., Mandel, J.L. and Chambon, P. (1980) Cell 20, 625-637.
3. LeMeur, M., Glanville, N., Mandel, J.L., Gerlinger, P., Palmiter, R.D. and Chambon, P. (1981) Cell 23, 561-571.
4. Maxam, A. and Gilbert, W. (1980) in Methods in Enzymology (L. Grossman and K. Moldave, Eds.) Vol. 65, pp. 499-560, Academic Press, New York.
5. Heilig, R., Muraskowsky, R. and Mandel, J.L., (1982) J. Mol. Biol. 156, 1-19.
6. Woo, S.L.C., Beattie, W.G., Catterall, J.F., Dugaiczyk, A., Staden, R., Brownlee, G.G. and O'Malley, B.W., (1981) Biochemistry 20, 6437-6446.
7. Breathnach, R. and Chambon, P.,(1981) in Ann. Rev. Biochem. (E.E. Snell, P.D. Boyer, A. Meister and C.C. Richardson Eds.) Vol. 50, pp. 349-383, Annual Reviews Inc..
8. Mount, S.M., (1982) Nucleic Acids Res. 10, 459-472.
9. Proudfoot, N.J. and Brownlee, G.G. (1976) Nature 263, 211-214.
10. Benoist, C., O'Hare, K., Breathnach, R. and Chambon, P. (1980) Nucleic Acids Res. 8, 127-142.
11. O'Hare, K., Breathnach, R., Benoist, C. and Chambon, P. (1979) Nucleic

Acids Res. 7, 321-334.
12. Nisbet, A.D., Saundry, R.H., Moir, A.J.G., Fothergill, L.A., Fothergill, J.E. (1981) Eur. J. Biochem. 115, 335-345.
13. Lingappa, V.R., Lingappa, J.R. and Blobel, G. (1979) Nature 281, 117-121.
14. Shapiro, H.S. (1976) in Handbook of Biochemistry and Molecular Biology, Nucleic Acids, (G.D. Fasman, Ed.) Volume II, 3rd Edition, p. 269, CRC Press Inc..
15. Bird, A.P. (1980) Nucleic Acids Res. 8, 1499-1504.
16. Staden, R. (1977) Nucleic Acids Res. 4, 4037-4051.
17. Westaway, D. and Williamson, R. (1981) Nucleic Acids Res. 9, 1777-1788.
18. Spritz, R.A., Jagadeeswaran, P., Choudary, P.V., Biro, P.A., Elder, J.T., DeRiel, J.K., Manley, J.L., Gefter, M.L., Forget, B.G. and Weissman, S.M. (1981) Proc. Natl. Acad. Sci. USA 78, 2455-2459.
19. Busslinger, M., Moschonas, N. and Flavell, R.A. (1981) Cell 27, 289-298.
20. Hagenbüchle, O, Bovey, R. and Young, R.A. (1980) Cell 21, 179-187.
21. Jung, A., Sippel, A.E., Grez, M. and Schütz, G. (1980) Proc. Natl. Acad. Sci. USA 77, 5759-5763.
22. Henderson, J.Y., Moir, A.J.G., Fothergill, L.A. and Fothergill, J.E. (1981) Eur. J. Biochem. 114, 439-450.
23. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. and Dodgson, J. (1980) Cell 20, 555-566.
24. Jeffreys, A.J. (1981) in Genetic Engineering (R. Williamson, ed.) Vol. 2, pp. 1-48, Academic Press, London.
25. Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.L., Blechl, A.E., Smithies, O., Baralle, F.E., Shoulders, C.C. and Proudfoot, N.J. (1980) Cell 21, 653-668.
26. Smithies, O., Engels, W.R., Devereux, J.R., Slightom, J.L. and Shen, S. H. (1981) Cell 26, 345-353.
27. Setlow, P. (1976) in Handbook of Biochemistry and Molecular Biology, Nucleic Acids, (G.D. Fasman, Ed.) Volume II, 3rd Edition, pp. 316-318, CRC Press Inc..
28. Slightom,J.L., Blechl, A.E. and Smithies, O. (1980) Cell 21, 627-638.
29. Shen, S.H., Slightom, J.L. and Smithies, O. (1981) Cell 26, 191-203.
30. Kunkel, G.R. and Martinson, H.G. (1981) Nucleic Acids Res. 9, 6869-6888.
31. Strauss, F., Gaillard, C. and Prunell, A. (1981) Eur. J. Biochem. 118, 215-222
32. Moreau, J., Matyash-Smirniaguina, L. and Scherrer, K. (1981) Proc. Natl. Acad. Sci. USA 78, 1341-1345.