Volume 14 Number 24 1986 Nucleic Acids Research

Compilation and analysis of eukaryotic POL II promoter sequences

Philipp Bucher and Edward N.Trifonov

Department of Polymer Research, The Weizmann Institute of Science, Rehovot, PO Box 26, Israel

Received 20 August 1986; Revised and Accepted 30 October 1986

ABSTRACT

A representative set of 168 eukaryotic POL II promoters has been compiled from the EMBL library and subjected to computer signal search analysis. Application of this technique to *E. coli* promoters as a control ensemble revealed the well known consensus sequences at -35 and -10 which indicates that the methods are adequate to approach problems of this kind. The results obtained from the eukaryotic promoter set can be summarized as follows: (i) Common sequence features are confined to a region between -50 and +10 relative to the transcriptional initiation site. (ii) The only well conserved consensus sequence is TATAAA, centered at -28. (iii) A weak motif, CA followed preferentially by pyrimidines, surrounds the cap-site. (iv) Two pentanucleotides which have been shown by experiments to stimulate transcription of certain genes, GGGCG and CCAAT, are moderately over-represented in the upstream region (between -129 and -50). However, they occur at highly variable distances from the initiation site.

INTRODUCTION

Eukaryotic POL II promoters have been the subject of intense investigation during the last decade. Despite these efforts, no generally accepted description of their general sequence features, such as exists for E. coli promoters, has as yet emerged. The results of earlier comparative studies (1,2) were derived from relatively small promoter sets biased by high proportions of histone and globin sequences and need re-evaluation. Site-directed mutagenesis data do not provide a coherent picture of promoter structure because it is usually not possible to decide whether the mutations affect general or gene-specific mechanisms. The only undisputed eukaryotic POL II promoter element is the Goldberg/Hogness- or TATA-box (3) which occurs between 25 and 30 bp upstream from the initiation site. Its requirement for accurate initiation as well as for maximal rate of transcription has been demonstrated for a considerable number of genes. However, in some cases it has also been shown that it is dispensable for low levels of transcription (4) or insufficient for high rates (5).

Many mutations which modulate the activity of a promoter have been mapped to a region upstream from the TATA-box (6). What remains uncertain is whether a second universal promoter element exists in this region which is inactivated by some of these mutations. Several candidate consensus sequences have been proposed for this. The most popular one is the CAAT-box introduced in two different versions by Efstratiadis *et al.* (7) and Benoist *et al.* (8). Although its quality as a consensus sequence has never been convincingly demonstrated by

comparative DNA sequence analysis, the biological significance of this motif seems to be broadly accepted. This is reflected by dozens of underlined or boxed CAAT-related oligonucleotides in newly published promoter sequences which are declared as "transcription signals" in the absence of experimental evidence that would support such a claim. The clarification of the status of this consensus sequence was one of the objectives of this investigation.

In our sequence analysis strategy we considered it as important to apply the following principles: We required first that the data set is representative in the sense that it does not include significant numbers of sequences which are closely related by phylogeny, and second that the algorithms do not depend on *a priori* assumptions on the nature of the sequence features to be found. Since we felt that a possible negative result would not be appreciated unless the power of our methods is demonstrated on a related problem where there is agreement about the expected results, we applied our sequence analysis procedures simultaneously to Hawley and McClure's collection of *E. coli* promoters (9) and show these results here, too.

SELECTION OF DNA SEQUENCE DATA:

We define eukaryotic promoters as DNA segments which determine the site rather than the rate of transcriptional initiation. The existence of transcriptional enhancers which influence initiation rates over distances of 1 kb or more renders alternative definitions impractical. Our compilation is therefore a collection of transcriptional initiation sites. Consequently we considered only biochemical but not genetic evidence in order to decide whether a given sequence should be incorporated or not. We further assumed that all capped 5'termini of eukaryotic mRNAs are generated by RNA POL II initiation. Biochemical evidence for a transcription start site usually comes from direct or indirect sequence analysis of mRNA 5'regions. In a few cases, data on the structure of *in vitro* generated transcripts were also accepted as promoter definition. Some cap-sites were inferred from experimentally determined transcriptional intiation sites of closely related genes. Putative promoters predicted from nucleotide sequence alone are not included in our compilation. However, in order to avoid subjective decisions, we did not exclude initiation sites located at unusual distances from a clear TATA-box if they were reportedly mapped by adequate techniques.

For purely technical reasons we confined our collection to sequences which were available in the EMBL nucleotide sequence data library release 7 (10). Promoters from lower eukaryotes (protozoa, slime-molds, algae, and fungi) were excluded because there are some indications that the specificity of their POL II transcription system might differ from that of higher eukaryotes. In an *in vitro* study, RNA polymerase II from yeast behaved more like E. coli polymerase than like the corresponding enzyme of higher eukaryotes (11). This taxonomic selection criterion applies only to the organisms where a given gene is expressed but not to the species which it belongs to due to its way of perpetuation. Consequently, our compilation includes many viral promoters as well as a few transcriptional initiation sites on the TDNA of Ti-plasmids, a DNA

segment which is replicated in a prokaryote but expressed by plant tumor cells after transformation (12).

Since the objective was to compile a set of promoters which is representative of higher eukaryotic genes in general, we had to eliminate a certain number of sequences which are closely related by phylogeny to other items of the collection. In doing so, we gave preference to the representatives with the longest upstream sequences available. The threshold for exclusion was set at 50% average homology between positions -50 and +10 relative to the initiation site. In principle, our sequence collection should also be devoid of larger groups of co-ordinately regulated promoters which could introduce statistically significant numbers of control signals into the ensemble which then could not be distinguished from general promoter elements by our computer analyses. With hemoglobin promoters constituting the largest subclass of this type but accounting only for 5% of the sequences in our compilation, we decided that further exclusions were not necessary.

Our computer algorithms require an initial alignment of the sequences with respect to an experimentally determined position. The fact that most transcriptional initiation sites are not mapped with absolute precision poses no fundamental problems for our techniques. However, difficulties arise when alternative transcription start sites are shown or supposed to be used by RNA polymerase for transcription of the same gene. In such a situation, we distinguished three cases. If most mRNA termini map to a small DNA region less than 10 bp in length, the sequence is listed only once in the collection and aligned with respect to an averaged position. If two or a few well separated major transcription start sites exist which are of similar strength or differentially regulated, each one appears as a distinct item in our compilation. If the pattern of transcriptional initiation is too diffuse to meet either of these conditions, the promoter was excluded from our set. Only a maize zein gene (13) and the late promoter region of polyoma virus were (14) discarded for this reason.

The analysis of E. coli promoters was based on Hawley and McClure's compilation (9). Only the precisely mapped promoters listed in Table 1 were considered. Those which were found in the EMBL library (85 out of 112) were analysed further upstream and downstream from the sequence segments shown in the original compilation.

COMPUTER METHODS FOR DNA SEQUENCE ANALYSIS

All analyses were carried out with an extended version of the signal search analysis program package described in detail by Bucher and Bryan (15). This method has much in common with Waterman's recently published pattern recognition techniques (16,17) and the package resembles in its software design certain parts of the "Delila system tools" described by Schneider *et al.* (18). A typical signal search analysis involves the following steps: 1. A set of fixed length DNA sequence segments defined by their location relative to an experimentally determined functional site (in this case transcriptional initiation sites) is extracted from a data

Gene and organism	−40 −30 −20 −10 0 +10
Wheat H3	TCTCGGTGCTCCTCCTATTTAACTCCGCCCCGTCCCCTTCTTCCTCCTCACCCCAATCTC
Wheat H4	CAACCTCTCGACCCCTTTAAGACGCCCTTCGCCCCACCCAGCAAATCACAGGCACCAGACG
Maize zein zA1	AATATTTGAGACCTCACCTATATAAATAGCTCCCATATCAGTAGTTAATCCATCACCCAT
Maize zein 19K	CACAAGGACTGAGATGTGTATAAATATCTCTTAGATTAGCTAGCTAATATATCGCACATA
Soybean RuBCC SS	ACACAAATCGACACTATTATATATAGCAAGTTTGAGCAGAAGCTTGGATATCTGGCAGCA
Soybean Lb I	CTCTTCAAGCCTTCTATATAAATAAGTATTGGATGTGAAGTTGTTGCATAACTTGCATTG
Soybean hs6871	TATATTGCTCCTCTACATCATTTTAAATACCCCATGTGTCCTTTGAAGACACATCACAGA
Soybean Le1	AAGTACCCAATAATGCTAGTATAAATAGGGGCATGACTCCCCATGCATCACAGTGCAATT
F.v. phaesolin	CTCTCTTATATAATACCTATAAATACCTCTAATATCACTCACTTCTTTCATCATCCATCC
A.t. TDNAo tmr P1	AATGAATTTCAAGGAGACAATATAACCGCCTCTGATAACACAATTCTCTAATATAAAAAT
A.t. TDNAo tmr P2	CTGATAACACAATTCTCTAATATAAAAATCAGTTTGTATTCAATATACTGCAAAAAACTT
A.t. TDNAo ocs	TTGCCCATTCATTGATCTATTTAAAGGTGTGGCCTCAAGGATAATCGCCAAACCATTATA
A.t. TDNAn nos	CAAAAATGCTCCACTGACGTTCCATAAATTCCCCTCGGTATCCAATTAGAGTCTCATATT
A.t. TDNAo tr-7	CGTCCCAGCCCGGCATCTATATATAGCGCCAATATAGTTTGTCTTACACAAACACACCTC
CAMV 8s, 35s-major	TTCGCAAGACCCTTCCTCTATATAAGGAAGTTCATTTCATTTGGAGAGGACACGCTGAAA
CAMV 35s-minor	TAATGACCCCTTATCGATTTAAAGAAATAATCCGCATAAGCCCCCGCTTAAAAAATTGGT
D.m. cut.-protein I	ACTTTGGGTGGCAATCATATAAAAAGGCTCTGCCCGACCACAATCAGTTATCAGTCAACG
D.m. cut.-protein III	TGCATCAGCTTTTGATGATATATAAACACCGATTTGAGCATAGATTGTCATCAGTCTTAG
D.m. sgs4 glue	CGATGGCAAAATGCAGTGGGTATATAAAGAGCATCAAGCGGTATTGAATTCCAAAGTCAA
D.m. 3L 74F	TATGTAATCATATAGATTCTATAATAAACAAAGAAACAAAACTAGTTGTAAAACAAACAC
D.m. globin IV	TTCTCAAAATTTTTAAGTATAAATGGAGCACAAATTTCGATAGTAAATCAGTTCTTCAAT
D.m. YP I	CGCTCAGCGTAAATTGTGGTATATAAACCACCATCGTTGGATTTGGAAGGCCAGTTCAAC
D.m. YP II	ATAGACTACCGATTCCAAGGGGTATAAAATGCATTGAGTCGCAGCAGTGGGCATGCAGTA
D.m. ADH larval	TGCTGTACGGATCTTCCTATAAATACGGGGCCGACACGAACTGGAAACCAACAACTAACG
D.m. ADH adult	CCCCCACGAGAGAACAGTATTTAAGGAGCTGCGAAGGTCCAAGTCACCGATTATTGTCTC
D.m. hsp 70K	CGAAAAGAGCGCCGGAGTATAAATAGAGGCGCTTCGTCGACGGAGCGTCAATTCAATTCA
D.m. hsp 22K	TTCCTCTCTGTCAAGAGTATAAATAGCCACCGGTTGGACACTACGCTCTCAGTTCAAAAA
D.m. hsp 23K	TTCGACAGCAAGCGGTTGTATAAATATCCGGCACTTTCGTGCAACCGGCGTCAGTTGAAT
D.m. hsp 26K	AGAAAAGCTCCAGCGGGTATAAAAGCAGCGTCGCTTGACGAACAGAGCACAGATCGAATT
D.m. hsp 27K	TGTGAGCCCAGCGTCAGTATAAAAGCCGGCGTCAACGTCGCCCGAGCACAGTCTAAACTG
D.m. hsp 68K	TCCCCTCCCGGCGACAGAGTATAAATACGGGCGCAAATTTCCCAGACGCTACATTTGAAA
D.m. hsp 83K	TTCGGGTGCGGGTTTTTCTATAAAAGCAGACGCGCGGCGTTTGCCGGTTCGAGTCTTGAA
D.m. 44D gene H	CACCTTATCGACTAGTATAAAAGGCACTGTCAGCTCTCCAGCCCGAACAAAATCGATCAA
D.m. 44D gene L	CAATGGGAGCGGTATGCTTAAATAGGGGCACCTTTTAATCCCTCTGGCCATTGGCAATCG
D.m. rp49	TATTTCCAGTGGGTCAGTGCACTAATGGCTACACTTGTTGTGTCCTACCAGCTTCAAGAT
B.m. fibroin	AAAACTCGAAAATTTTCAGTATAAAAAGGTTCAACTTTTTCAAATCAGCATCAGTTCGGT
B.m. Hc-A.13	GGTGAACATGATTCTTAGTTACTATATAAGAACGAAGTCTTAAGCTTTAAGTATTCAAGA
B.m. Hc-B.13	ATTTTCAAGGAAACTGCTCGGTATAAAAGCTGATGTAGTTCAGAGTTAAGTCATTCTGAA
P.m. early H1	CCACGTACGCAACCGCGCGGGATATAGGTGAGGTTGCCGTGAGGGCCGTCACTTGTTTTG
P.m. early H2A	TCCGATCCCGACGTTTGGTATAAATAGCCAGCAAAAAAGATAGGTGGTCAACCATTCAAG
S.p. early H2B	ACGGATCCGGCCCCGTGTATAAAAAGGAAAGGTTCTCGCTGGCCATTCACAGTATCCAAA
S.p. early H3	CCAGGATCCCGCAGCACATATAAATAGCTGAAAATTGCCAGTGGTTCTCATTCATCCCGT
S.p. early H4	CAAGTCCGCAATGGTGTAACAATACTCGGTGCAATCCGGTTGAGGCATCATTCGCTTAGC
L.p. late H3	CGAGAAGCAGTCTGGAGGTATAAATACGTCGCGGTTACTTTGAAAATTTATCAGTTGACT
L.p. late H4	TAAAGGCTATATATACCGCACGAACAGCAGAATTGAGTATCAGTTTGAATCTCAAACAGG

Exp. def.	Expression/Regulation	References for initiation site	EMBL Sequence		Ref.
3	proliferating tissues	MGG196:397	TAHI02	1+	186
3	proliferating tissues	NAR11:5865	TAHI01	1+	669
3	endosperm	EMBOJ1:1589	ZMZE05	1+	148
4	endosperm	Cell29:1015	ZMZE01	1+	888
3	leaves, +light	JMAG1:483	GMRUBP	1+	241
3	root nodules	PNAS79:4055	GMGL04	1+	144
3	root *e.g.*, +heatshock	EMBOJ3:2491	GMHSP2	1+	492
4	cotyledon	Cell34:1023	GMLEA	1+	942
4,6	cotyledon	PNAS80:1897	PVPHASL	1+	101
3	plant tumor	NAR11:6211,JMAG2:354	ATACH5	1+	8729
3	plant tumor	NAR11:6211,JMAG2:354	ATACH5	1+	8760
3	plant tumor	JMAG1:499	ATACH5	1-	13658
3	plant tumor	NAR11:369,JMAG1:561	ATNOPA	1+	550
3	plant tumor	EMBOJ2:419	ATACH5	1-	3303
3,8	infected leaves	Cell30:763	CAMVG2	0+	7435
3,8	infected leaves	Cell30:763	CAMVG2	0+	8017
6	third instar larva	Cell29:1027	DMCUT1	1-	760
6	third instar larva	Cell29:1027	DMCUT2	1+	2606
3,7	larva; salivary glands	Cell29:1041,Cell34:74	DMSGS4	1+	52
4	larva; salivary glands	EMBOJ3:289	DM74EF	1+	401
4	larva; fat body	Nature310:795	CTGL01	1+	260
3,7	puppa; ovary, fat body	NAR10:2261	DMYOLK1	1-	225
3,7	puppa; ovary, fat body	NAR10:2261	DMYOLK1	1+	1447
4,5	larva; fat body, gut	Cell33:125	DMADH1	1+	974
4,5	adult	Cell33:125	DMADH1	1+	267
4,5	+heatshock	NAR8:3105,Cell21:669,EMBOJ1:1583	DMHSP1	1+	717
4,8	+heatshock	NAR9:1627	DMHS08	1+	514
4,8	+heatshock	NAR9:1627	DMHS09	1+	320
3,8	+heatshock	NAR9:1627,PNAS78:3775	DMHS10	1+	470
4,8	+heatshock	NAR9:1627	DMHS11	1+	290
3	+heatshock	PNAS78:3775	DMHSP68	1+	158
3	+heatshock	NAR11:7011,PNAS78:3775	DMHS83	1+	878
3	larva, adult	JMB166:101	DMCUT3	1-	3169
3	larva, adult	JMB166:101	DMCUT3	1-	9158
3	housekeeping gene	NAR12:5495	DMRP49	1+	411
1,3,6	larva; silk gland	Cell16:425,Cell18:591	BMFIBR	1+	551
(3 or 4)	eggshell, late	PNAS81:4452,JME20:265	BMCH01	1-	248
(3 or 4)	eggshell, late	PNAS81:4452,JME20:265	BMCH01	1+	514
3	early blastula	Nature285:147,Nature288:100	PMHIS7	0+	4860
3	early blastula	Nature285:147,Nature288:100	PMHIS7	0+	3614
5	early blastula	Nature279:737,PNAS77:1265	SPHIS1	1+	170
5	early blastula	PNAS77:1265	SPHIS1	1+	1341
1,5	early blastula	Bioch20:1216,PNAS77:1265	SPHIH4	1+	165
3	late blastula	Cell31:383,PNAS81:2411	LPHISL34	1+	1487
3	late blastula	Cell31:383,PNAS81:2411	LPHISL34	1-	724

Gene and organism	−40	−30	−20	−10	0	+10
Trout protamine	ACTCCAGCCCCCTCCAGCCCTATAAAAGGGAGCACGGCCGTCTAAAAGTCTTATCCATCA					
Chicken H1	TCACCGCGCGGCTCCGCTCTATAAATACGAGGCCGCCGACTTGCTCCGGGCCCAGTGGTT					
Trout testis H2A	CAGACGCCGCTGCCGGCCTTATAAACTTCACATAGGCATTTTGAGGCTATACTCCGACTG					
Trout testis H3	GGCTTTTGTGGCGAGGTATAAGTAAGGCTCTCGAGGTGCCCAGCGGCTCATTCAGACTTT					
Chicken H4	GGTCCGACCATACGCCATAACACCCGCGCGCGCCCCGCCACATCCTCACTGGTGTCGGAC					
Xenopus H4	CAGGTCCTCTCCAGCTGCATATAAAGAGGAGGAGAGGCCCTGATACGTTATATTGTGTTT					
Mouse H4	TCTGGTCCGATCCTCTCATATATTAGTGGCACTCCACCTCCAATGCCTCACCAGCTGGTG					
Human SOD-1	GCGAGGCGCGGAGGTCTGGCCTATAAAGTAGTCGCGGAGACGGGGTGCTGGTTTGCGTCG					
Mouse MT-I	CGCCCGGACTCGTCCAACGACTATAAAGAGGGCAGGCTGTCCTCTAAGCGTCACCACGAC					
Human MT-IIA	TCGTCCCGGCTCTTTCTAGCTATAAACACTGCTTGCCGCGCTGCACTCCACCACGCCTCC					
Human DHFR	GGGGGCGGGGCCTCGCCTGCACAAATAGGGACGAGGGGGCGGGGCGGCCACAATTTCGCG					
Mouse DHFR	GCCTAAGCTGCGCAAGTGGTACACAGCTCAGGGCTGCGATTTCGCGCCAAACTTGACGGC					
Mouse HPRT	CGAGAGGGCGGGCCGAGGGGCGGAGCCTGGCCGGCAGCGTTTCTGAGCCATTGCTGAGGC					
Chicken α-actin	GGCCGGGCGGTGCTCCCGTCGATAAAAGGCTCCGGGGCCGGCGGGCGACCCGAGCTACCC					
Rat skel. muscle actin	TGGAGAGCTCAGGACTATATAAAAACCTGAGGCTAGGGACAGGCGGTCACACGGACGTGA					
Chicken β-actin	GAGGCGGCGGCGGCGGCGGCCCTATAAAAAGCGAAGCGCGCGGCGGGCGGGAGTCGCTGC					
Rat β-actin	CGAGTGGCCGCTGTGGCGTCCTATAAAACCCGGCGGCGCAACGCGCGCCACTGTCGAGTC					
Chicken myosin LC1	TGTACAAGGCGCTAAGTGAAATATATATATGCCCTTATAGAGTTTAGCACACTGGGTCCA					
Chicken myosin LC3	CAGCAATGCCGTCGCGCTGCCAGATAAATAAGGGGAAGAAAGGCCAGGAAAGCAGGACCA					
Mouse myosin LC2	GGTATGTTAAGGGGCAGGACTATATAACCCCAGAAGAACTGCCCCAAGCAGATTCTCTGC					
Chick. α2(I)-collagen	GCGGGACCCCCTGCGGTATAAATACGGCGGAGCGGGGCTTGATTAATTTAGCATCCCGGG					
Mouse α1(I)-collagen	TCCCAGCTCTCCATCAAGATGGTATAAAAGGGGCCCAGCCAGTCGTCGGAGCAGACGGGA					
Chicken f.-keratin	GCCTACTATAGTTACATATGCATAAATTAACTCTAAACCAGGCTCCCTCATCCACTTCTC					
Mouse β-crystallin	ATCCTGGGTTGTAGCTAGTTATGAAAACCACAGGATGAAGTTTGTTCTTAACTTGCACCC					
Seal myoglobin	GTCAAGCTTCTGGGAAAGTATAAAATCCCTCTGGGGCCAGGCGATCTCAAACCCCAGCTG					
Human α-globin	GCGTGCCCCCGCGCCCCAAGCATAAACCCTGGCGCGCTCGCGGCCCGGCACTCTTCTGGT					
Mouse α-globin	AGGACAGCCCTTGGAGGGCATATAAGTGCTACTTGCTGCAGGTCCAAGACACTTCTGATT					
Rabbit β-globin	CATAGTTCAGGACTTGGGCATAAAAGGCAGAGCAGGGCAGCTGCTGCTTACACTTGCTTT					
Rabbit β3-globin	AGATGTCCAGCGAGGAAGAATAAAAGGACGAGCCTTAGAGCAGTTTCACATACTTGCTTC					
Chicken β-globin	GGAGGGGCCCGGCGGAGGCGATAAAAGTGGGGACACAGACGGCCGCTCACCAGCGTGCTA					
Chicken ε-globin	GAGGAGCTGTCAGCGGTGGATAAAAGCCCCGGGGGTCCGCAGCTCCGCTCCAAGCTCTGA					
Human γA-globin	GGCTGGCTAGGGATGAAGAATAAAAGGAAGCACCCTTCAGCAGTTCCACACACTCGCTTC					
Xenopus βI-globin	TGACTCAGCATGGCCATATAAAGCAAGGCCAACAACTCAAAGGAACAGCAGCCTCTTACT					
Rat TAT	ACGCCCATTGGCTGAAACTATTTCAAGGGTCAGGACTGCACCTGAGCTCATCATCAGAGG					
Rat liver p-450	CTGAGTGTAGGGGCAGATTCAGCATAAAAGATCCTGCTGGAGAGCATGCACTGAAGTCTA					
Chicken serum alb.	AAGCAGTCAGTAAAAGGTATATAAGAAAATGATTTCCCTCAATCATCCTAGCATTTTTGA					
Chicken ovalbumin	GTGGGTCACAATTCAGGCTATATATTCCCCAGGGCTCAGCCAGTGTCTGTACATACAGCT					
Chicken gene X	GTGTCCGAAAGGGTACTGTATATATCACCAAGGACTCAGAGAATCTGTTCAGGTTCAACT					
Chicken gene Y	TGTCATGACATTATACAGGATATATTTCAAGGAGTTCTGCAAGGCTGTACCACGTACAGC					
Chicken conalbumin	CAGCCAGGGCTGCTCCTCTATAAAAGGGGAAGAAAGAGGCTCCGCAGCCATCACAGACCC					
Chicken ovomucoid	----------TTTGTATATATTTGCAGGCAGCCTCGGGGGGACCATCTCAGGAGC					
Chicken lysozyme	AAAGGGGGTGGGAGGAAGTTAAAAGAAGAGGCAGGTGCAAGAGAGCTTGCAGTCCCGCTG					
Xenopus vitellogenin	GTGTTACAGATTTTCCTGCAATAAATATGGCAGGCTTTTCTGGGTTCAGTGTTCACCATC					
Chicken VTGII	GTTCCTGAACATTCTTCCATAAAAGTCTCACCATGCCTGGCAGAGCCCTATTCACCTTCG					
Chicken apoVLDLII	CCCTCACTATATTAGTTCTGCATAAATGCCAGTGTCTCAGATGAGCATCAACCTCAGCTT					

Exp. def.	Expression/Regulation	References for initiation site	EMBL Sequence	Ref.
(3,6)	spermatocytes	NAR10:7581,NAR10:4551,NAR11:4907	SGPROTA1 1+	252
3	embryo	JBC258:9005	GGH11A1 1+	167
3	spermatogones	JME20:236	SGHIS2A3 1+	1192
3	spermatogones	JME20:236	SGHIS2A3 1+	329
3	embryo	JBC258:9005	GGH43D8 1+	244
4°	not active in oocytes	NAR11:8641	XLHIS4 1+	380
3	during S-phase	JMB151:607,Cell41:885	MMHI01 1+	229
3,8	housekeeping gene	NAR12:9349	HSSOD1G1 1+	292
3	+heavy metal ions	Nature292:267	MMMTIX 1+	301
3	+heavy metal ions	Nature299:797	HSTHI02A 1+	300
2*,3‡,8‡	cell cycle: G1/S	JBC259:3933	HSDHFR01 1+	324
4,8	cell cycle: G1/S	JBC261:4685,MCB6:365	MMDHF5 1+	388
4,8	housekeeping gene	PNAS81:2147,Cell44:319	MMHPRT1 1+	846
4,5	embryo; skeletal muscle	NAR10:3861	GGACTI 1+	92
3	skeletal muscle	Nature298:857	RNAC02 1+	193
7	housekeeping gene	NAR11:8287	GGAC01 1+	544
3	housekeeping gene	NAR11:1759	RNAC01 1+	235
3	skeletal muscle	Nature308:333	GGMY03 1+	321
3	skeletal muscle	Nature308:333	GGMY04 1+	344
4	skeletal muscle	NAR12:7175	RNMYOLC1 1+	237
1*,3,6	embryo; fibroblasts	JBC256:11251,PNAS78:5334	GGC1A201 1+	404
3	foetus	PNAS81:1504,Nature304:315	MMC1A1LV 1+	220
5	embryo; feather	NAR10:6007	GGKERC 1+	61
3	lens	Nature302:310	MMCRY1 1+	71
4	skeletal muscle	Nature301:732	HGGL01 1+	262
1,6	adult; reticulocytes	JBC255:2807,Cell12:1085	HSAGL1 1+	98
1,4‡	adult; reticulocytes	JBC252:1758,Cell21:697	MMAGL1 1+	372
1	adult; reticulocytes	Cell9:747,Cell32:695	OCBGLO 1+	224
3	embryo; reticulocytes	JBC256:11780	OCBGLX 1+	162
3,7	adult; reticulocytes	JBC258:3983,Biochi20:2091	GGGL02 1+	386
4,6	embryo; reticulocytes	JBC258:12685,Cell28:515	GGHBBR2 1+	199
6	foetus; reticulocytes	NAR5:3515	HSGLBN 1+	7062
4	larva; reticulocytes	NAR12:7705	XLBGL3 1+	241
3	liver, +glucocorticoid	PNAS81:1346	RNTAT5E 1+	601
3	liver, +phenobarbital	PNAS80:3958	RNCYP451 1+	71
4	liver	JBC258:4556	GGAL07 1+	267
1	oviduct, +estrogen	NAR9:1657	GGOV03 1+	1342
3	oviduct, +estrogen	JMB156:1	GGOV01 1+	1327
3	oviduct, +estrogen	JMB156:1	GGOV02 1+	1612
3,5	oviduct, +estrogen	Nature282:567	GGCALB1 1+	267
3,6	oviduct, +estrogen	JCB87:480,JMB162:345	GGOV01 1+	35
3	oviduct, +estrogen	Cell25:743	GGLYSX 1+	439
3,5	liver, +estrogen	EMBOJ2:2271	XLVITE 1+	494
4,5	liver, +estrogen	EMBOJ2:2271,NAR12:1117	GGVI01 1+	1146
2*,3	liver, +estrogen	NAR11:2529,JBC258:4556	GGVL01 1+	485

Gene and organism	-40 -30 -20 -10 0 +10
Rat α-lactalbumin	GTGCTAGGGCCAGAGGCCTTCTTCATAAATAAAAGCAGGTGAAGTGAGTGGGATCCACAT
Rat γ-casein	GATGCTAGAACCTGGTTTAAATAGTGCGGGAGCTACCCACTGCTATCATCATCTACCTAT
Mouse complement C3	GGACCAGAGAGGAGAGCCATATAAAGAGCCAGCGGCACAGCCCCAGCTCGCCTCTGCCCA
Rat γ-fibroin	CCCGCCCAGACTGGGAATTCATATAAAGGCCCAAGGAGAGCCCAAGAGGTCACAGTGCTG
Human factor IX	CAGAAGTAAATACAGCTCAGCTTGTACTTTGGTACAACTAATCGACCTTACCACTTTCAC
Mouse kallikr. mGK-1	CTGTGGGGAGAATGGGGGATTTAAAGTCTCCCCAGGGAGCCTCAATAGCTCCAAGCTCAC
Mouse α-amylase	AATGTACTTTTTGTAGAAATATAAATAGGCGCTAGAGAGAAAGAACACTGACAACTTCAA
Rat PSBP C3	AGGTGATTGCCTGAGCAATAAATAGAGGAACACTGAGGTCTCAGCTCCAGAGTTTCCTGA
Rabbit uteroglobin	GGGCACTGCCCGGAGAATACAAAAAGGCACCTGACGGCCGTCCCCCTCAAGATCACCGGA
Rat vasopressin	TCCTAGCCAACACCTGCAGACATAAATAGACAGCCCAGCCCGCTCAGGCAGCAGAGCAGA
Rat oxytocin	CCCACCATGGCAGTGGACAAGGCATAAAAAGGTCGGTCTGGGCTGGAGAAACCATCACCG
Bovine oxytocin	CGCCCACGCGGCCGCCGGGCTTAAAAGGCCAGACCCGAGAGACGGCCGCAGTCCCCGGCC
Bovine prolactin	ATTCATGAAGATGTCAAAGCCTTATAAAGCCAACATCTGGGGAAGAGAAAGCCATAGGAC
Rat growth hormone	TCGAGGAAAACAGGTAGGGTATAAAAAGGGCATGCAAGGGACCAAGTCCAGCACCCTCGA
Human ACTH/β-LPH	CCACCAGGAGAGCTCGGCAAGTATATAAGGACAGAGGAGCGCGGGACCAAGCGGCGGCGA
Hum. CG/LH/FSH/TSH	GGTGGAAACACTCTGCTGGTATAAAAGCAGGTGAGGACTTCATTAACTGCAGTTACTGAG
Human enkefalin A	TTCGGTTTGGGGCTAATTATAAAGTGGCTCCAGCAGCCGTTAAGCCCCGGGACGGCGAGG
Rat parath. hormone	GGCATGACATCATCCTTCCCAATAAAATACTCCTCTTGGTGAGCAAAAGGCCTGCATATG
Human insulin	GGGAGATGGGCTCTGAGACTATAAAGCCAGCGGGGGCCCAGCAGCCCTCAGCCCTCCAGG
Chicken insulin	----------CTTCTGGTTATAATTGGTCATTTATTATGACTTTTAAAGCCTGATGAA
Human α-interferon	GAAATTAGTATGTTCACTATTTAAGACCTATGCACAGAGCAAAGTCTTCAGAAAACCTAG
Human β-interferon	TAGAGAGAGGACCATCTCATATAAATAGGCCATACCCACGGAGAAAGGACATTCTAACTG
Human γ-interferon	CCTCAGGAGACTTCAATTAGGTATAAATACCAGCAGCCAGAGGAGGTGCAGCACATTGTT
Human IL-2 (TCGF)	AATATTTTTCCAGAATTAACAGTATAAATTGCATCTCTTGTTCAAGAGTTCCCTATCACT
Mouse Ig VH101	AAGCAGCCCTCAGGCAGAGGATAAAAGCTCACACTAACTGAGAAGCTCCATCCTCTTCTC
Mouse Ig V1	AATTAGGCCACCCTCATCACATGAAAACCAGCCCAGAGTGACTCTAGCAGTGGGATCCTG
Human Ig κ HK101	CTCCTGCCCTGAAGCCTTATTAATAGGCTGGTCAGACTTTGTGCAGGAATCAGACCCAGT
Mouse Ig κ T	TCACTGCCTTGGGGACTTCTTCATATACCCGTCACACATGTACGGTACCATTGTCATTGC
Mouse Ig κ MPC11	GCACTGAGGGCCAGCTGATTTATAAC-AGGTCTTTGCAGTGAGATATGAAATGCATCACA
Mouse Ig λI	CAGCCCAGCCCATACTAAGAGTTATATTATGTCTGTCTCACAGCCTGCTGCTGACCAATA
Human HLA-DR	TGCATTTTAATGGTCAGACTCTATTACACCCCACATTCTCTTTTCTTTTATTCTTGTCTG
Mouse MHCII Ia Ekα	AAAAGTTGAGTGCTTTGGATTTTAATCCCTTTTAGTTCTTGTTAATTCTGCCTCAGTCTG
M-MuLV LTR	GCTTCTGCTCCCCGAGCTCAATAAAAGAGCCCACAACCCCTCACTCGGGGCGCCAGTCCT
Human ATLV LTR	TCAATAAACTAGCAGGAGTCTATAAAAGCGTGGAGACAGTTCAGGAGGGGGGCTCGCATCT
Human ARV-2 LTR	TGGCGTCCCTCAGATGCTGCATATAAGCAGCTGCTTTTTGCCTGTACTGGGTCTCTCTGG
Avian RSV LTR	CCGCATCGCAGAGATATTGTATTTAAGTGCCTAGCTCGATACAATAAACGCCATTTTACC
Avian SNV LTR	ACCCTGTAAGCTGTAAGCGGCTATATAAGCCGGGTACATCTCTTGCTCGGGGTCGCCGTC
HSV-1 IE-I	TTTGGGGAGGGGAAAGGCGTGGGGTATAAGTTAGCCCTGGCCCGACAGTCTGGTCGCATT
HSV-1 IE-II	AGCCGGCCCCGGCACCACGGGTATAAGGACATCCACCACCCGGCCGGTGGTGGTGTGCAG
HSV-1 IE-III	TTCCCGCCGGCCCCTGGGACTATATGAGCCCGAGGACGCCCCGATCGTCCACACGGAGCG
HSV-1 IE-IV/V	GGGGGCGGGTCTCTCCGGCGCACATAAAGGCCCGGCGCGACCGACGCCCGCAGACGGCGC
HSV-1 early 33K	GGCCGGGCGACCCAGATGTTTACTTAAAAGGCGTGCCGTCCGCCGGCATGCACCCCAGAG
HSV-1 early 21K	CGACGTACGCGATGAGATCAATAAAAGGGGGCGTGAGGACCGGGAGGCGGCCAGAACCGC
HSV-1 early 5.0 kb	GCCCCACCCCTGCGCGATGTGGATAAAAAGCCAGCGCGGGTGGTTTGGGTACCACAGGTG
HSV-1 early 1.2 kb	TGGTCCGCCTTCTGGTCCACGCATATAAGCGCGGACTAAAAACAGGGATGTACTACTGCA
HSV-1 TK	CGCGGTCCCAGGTCCACTTCGCATATTAAGGTGACGCGTGTGGCCTCGAACACCGAGCGA
HSV-1 β/γ-late 6 kb	CGGACGCTTTGCCGCCTCTGCCAATTTCTTCCTGCACGCTTTTGGACCAGGGCCATCTTG

Exp. def.	Expression/Regulation	Referencesfor initiation site	EMBL Sequence	Ref.
3,6	mam. glands,+prolactin	Nature308:377,PNAS77:2093	RNLALBO1 1+	1248
6	mam. glands,+prolactin	NAR10:8079	RNCASG11 1+	96
7	liver, +hydrocortisone	PNAS79:7077	MMC31 1+	107
3	liver	Cell31:159	RNFBRG5E 1+	274
3,7	liver	EMBOJ3:1053	HSFIXG1 1+	296
8	submaxillary gland	Nature303:300	MMKALL 1+	4474
1,6	pancreas,	Cell21:179	MMAMY2 1+	434
4,5,6	prostata, +androgen	EMBOJ2:769,JBC258:12	RNPSO1 1+	584
4	+progesterone	PNAS79:4853	OCUG1 1+	396
3,(5)	hypothalamus	EMBOJ2:763,Nature295:299,EMBOJ3:3289	RNVNO3 1+	368
3	hypothalamus	PNAS81:2006	RNOXTNP 1+	220
3 or 4	hypothalamus	Nature308:554	BTHORO1 1+	210
6,(3)	pituitary	DNA3:237,JBC256:10524	BTPROLO1 1+	475
3,(5)	pituitary	NAR9:2087,NAR7:305,NAR9:3719	RNGROW3 1+	401
3	pituitary	EMBOJ1:1533,EJBC133:599	HSACTH 1+	681
8	placenta	JMAG1:3	HSAGC1 1+	92
3,8	adrenal medulla	EMBOJ2:2223 Nature297:431	HSENKE 1+	948
3,6	parathyroid gland	JBC259:3320	RNPTH2 1+	399
5,(4‡)	pancreas islet cells	Sci208:57,Nature306:557	HSINSU 1+	2186
4	pancreas islet cells	Cell20:555	GGINS1 1+	38
(3,6)	leukocytes, +viral inf.	Nature287:401,Sci212:1159	HSIFD1 1+	2194
3	fibroblasts, +viral inf.	PNAS78:5305	HSIFD4 1+	284
6	lymphocytes, +mitogen	NAR10:3605	HSIFNG 1+	347
8	T lymphocytes,+antigen	Nature302:305	HSIL05 1+	1366
3,6	B lymphocytes,+antigen	JBC257:277	MMIGHAI1 1+	237
4,5	B lymphocytes,+antigen	NAR10:7731	MMIGHAE 1+	575
3*,(5)	B lymphocytes,+antigen	NAR10:1841,Cell25:47	HSIGK2 1+	109
3	B lymphocytes,+antigen	EMBOJ1:719,Cell33:741	MMIG19 1+	840
1,3	B lymphocytes,+antigen	Cell29:681	MMIGKAL 1+	166
3,7 or 8	B lymphocytes,+antigen	PNAS80:417,EMBOJ4:2831	MMIG31 1+	221
3	lymphoid cells,+antigen	NAR11:8663	HSHL07 1+	449
3,7	lymphoid cells,+antigen	Cell32:745	MMMHO2 1+	94
1,3*,6	leukemia	Cell13:761,PNAS78:5411,PNAS77:3307	REMML1 1+	486
8	T-cell leukemia	PNAS79:6899	RE1PROP 1+	376
7	AIDS-inf. T-cells	Sci227:484	AIARV2 1+	455
1,6	sarcoma	Nature262:186,NAR10:5183,PNAS74:989	RERSV6 1+	9292
1,5 or 6	various cell-types	Nature285:550	REXXX1 1+	419
3	immediate early	JVIR44:939	HE1A0 1+	324
3	immediate early	NAR11:6271,JVIR43:1015	HE2IERN2 1+	269
3	immediate early	JGV62:1,PNAS79:4917,NAR11:2347	HE1IE3A 1+	371
3	immediate early	NAR10:2241,JGV62:1	HEHS08 1+	136
4	early	NAR12:2473	HEHS08 1+	1078
4	early	NAR12:2473	HEHS08 1+	784
3	early	PNAS78:6139,JGV64:997	HEHSV1 1+	121
3	early	JGV64:997	HEHSO6 1+	371
3,5	early	PNAS78:1441,NAR8:5949	HEHSTK 1+	407
3	intermediate/late	PNAS78:6139	HEHSV2 1+	111

Gene and organism	-40 -30 -20 -10 0 +10
EBV DL/DR region	ACAGAGACCCCAAAAAGAGGATAAAAGAAGGCGAGCCGGCCCGGCTCGCCAGCGTCGTCC
EBV BL-R1	GACAGGGACGGCGGCGCTATATATAAGAGCCCAAGACCCGGCTCTCTTTACTGCGAAATG
EBV BL-R2	CGGATTAGATGGGGATATTTAAAAGGGGCAGCAATCTCGGCTGTTTGTACTTCTTCTCTG
EBV BL-L2	ACCCAACAGGTGGTGAAAATATAACACAGGTGACACCAGCCTCTATCAGCACACATCATG
EBV BL-L1	ACCCCCCTTGTACCTATTAAAGAGGATGCTGCCTAGAAATCGGTGCCGAGACAATGGAGG
EBV BL-L3	CGGGTCTTGGGCTCTTATATATAGCGCCGCCGTCCCTGTCTGTTAGATCATCACCATGGA
EBV BK 2.1 kb	AGACGCCCTCAATCGTATTAAAAGCCGTGTATTCCCCCGCACTAAAGAATAAATCCCCAG
EBV BK 1.3 kb	TTGCGACCCCTCTGATATTAAGGTGGTTATTTTGGGCCAGGACCCCTATCACGGGGGTCA
EBV EH-L1	CGGTGCCCGGACTCAGAATTATTAAACCGGGTGGCAGCTCCTGGCAGTCATTCATTCGGA
EBV EC-L1	AAGGGCAGGGGGTGGGTATTTAAGGATCTATATGCCCTTCTCTACCTGCACCTCCAAATG
EBV ED-L1	CTCTGACGTAGCCGCCCTACATAAGCCTCTCACACTGCTCTGCCCCCTTCTTTCCTCAAC
Ad2 EIa	GTCAGCTGACGCGCAGTGTATTTATACCCGGTGAGTTCCTCAAGAGGCCACTCTTGAGTG
Ad2 EIb	GGGGCGGGGCTTAAAGGGTATATAATGCGCCGTGGGCTAATCTTGGTTACATCTGACCTC
Ad7 EIb	TTCTTGGGTGGGGTCTTGGATATATAAGTAGGAGCAGATCTGTGTGGTTAGCTCACAGCA
Ad12 EIb	TGGGCGTGGTTAAACAGGGATATAAAGCTGGGTTGGTGTTGCTTTGAATAGTTCATCTTA
Ad2 EII	GAAAGGGCGCGAAACTAGTCCTTAAGAGTCAGCGCGCAGTATTTGCTGAAGAGAGCCTCC
Ad2 EIII	TGCGGTCGCCCGGGCAGGGTATAACTCACCTGAAAATCAGAGGGCGAGGTATTCAGCTCA
Ad2 EIV	TTACGTCATTTTTTAGTCCTATATATACTCGCTCTGTACTTGGCCCTTTTTACACTGTGA
Ad2 IVa2	CCCTCCCACTTAGCCTCCTTCGTGCTGGCCTGGACGCGAGCCTTCGTCTCAGAGTGGTCC
Ad2 IX	GCTTAAGGGTGGGAAAGAATATATAAGGTGGGGGTCTCATGTAGTTTTGTATCTGTTTTG
Ad7 IX	ATGGGGACTTTCAGGTTGGTAAGGTGGACAAATTGGGTAAATTTTGTTAATTTCTGTCTT
Ad2 major late	GTGTTCCTGAAGGGGGGCTATAAAAGGGGGTGGGGGCGCGTTCGTCCTCACTCTCTTCCG
Ad2 LIIa	GGCGTGGTAGTCCTCAGGTACAAATTTGCGAAGGTAAGCCGACGTCCACAGCCCCGGAGT
AAV2 major mRNA	CCGCCCCCAGTGACGCAGATATAAGTGAGCCCAAACGGGTGCGCGAGTCAGTTGCGCAGC
AAV2 m.p. 0.06	CATGTGGTCACGCTGGGTATTTAAGCCCGAGTGAGCACGCAGGGTCTCCATTTTGAAGCG
AAV2 m.p. 0.19	GTGGACTAATATGGAACAGTATTTAAGCGCCTGTTTGAATCTCACGGAGCGTAAACGGTT
SV40 T/t antigen	TGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCGGCCTCTGAGCTATT
Polyoma T/t	GGCCACCCAAATTGATATAATTAAGCCCCAACCGCCTCTTCCCGCCTCATTTCAGCCTCA
SV40 T/t late	CCGCCCCTAACTCCGCCCAGTTCCGCCCATTCTCCGCCCCATGGCTGACTAATTTTTTTT
Polyoma T/t late	CTGTTTTTTTTAGTATTAAGCAGAGGCCGGGGACCCCTGGCCCGCTTACTCTGGAGAAAA
SV40 major late	GTTCTTTCCGCCTCAGAAGGTACCTAACCAAGTTCCTCTTTCAGAGGTTATTTCAGGCCA

Figure 1. Compilation of 168 eukaryotic POL II promoters. The sequences were selected according to the criteria described in the text. Underlined nucleotides correspond to capped 5'termini of mRNAs characterized by direct RNA sequencing. Dots point to regions where transcriptional initiation is likely to occur according to less precise mapping techniques. The numbers in the first column of the right-hand pages identify the experiments which define the promoter. They have the following meaning:

1 Direct RNA sequence analysis.
2 Length measurement of a transcript.
3 Length measurement of a nuclease-protected DNA fragment by comparison with a corresponding sequence ladder.
4 Length measurement of a nuclease-protected DNA fragment by comparison with unrelated molecular weight markers.
5 Indirect RNA sequencing by dideoxy-terminated cDNA synthesis.
6 DNA sequencing of an *in vitro* generated run-off cDNA or a full-length cDNA clone.
7 Length measurement of an *in vitro* generated run-off cDNA by comparison with a corresponding sequence ladder.
8 Length measurement of an *in vitro* generated run-off cDNA by comparison with unrelated molecular weight markers.

Exp. def.	Expression/Regulation	References for initiation site	EMBL Sequence Ref.
4	+TPA	JVIR56:987	EBV 1- 52787
2*,4	late	EMBOJ3:1083	EBV 1+ 88539
2*,4	late	EMBOJ3:1083	EBV 1+ 88897
4,8	early	EMBOJ3:1083	EBV 1- 90021
2*,4	late	EMBOJ3:1083	EBV 1- 92157
2*,4	early	EMBOJ3:1083	EBV 1- 88480
4,8	+TPA	JVIR54:501	EBV 1+109939
4,8	+TPA	JVIR54:501	EBV 1+110632
2*,4	+TPA	EMBOJ2:1331	EBV 1-137680
2*,4	+TPA	PNAS80:1565	EBV 1-159337
2*,4,7	latently infected cells	JVIR51:411,EMBOJ2:1331	EBV 1-169514
1	immediate early	JMB149:189,CSHSQB44:415	AD2 1+ 498
1	early, +E1a	JMB149:189,CSHSQB44:415	AD2 1+ 1700
3	early, +E1a	Gene18:143	AD7001 1+ 1577
3	early, +E1a	Cell27:121	AD1201 1+ 1527
1,3	early, +E1a	Cell18:569,JMB149:189,PNAS78:7383	AD2 1- 27092
1	early, +E1a	JMB149:189,CSHSQB44:415	AD2 1+ 27610
1	early, +E1a	NAR9:1675,JMB149:189	AD2 1- 35611
1,3,7	intermediate	JMB149:189,NAR10:7089	AD2 1- 5827
1	intermediate	JMB149:189,Cell19:671	AD2 1+ 3575
4	intermediate	Gene13:375	AD7001 1+ 3460
1	early/late, +E1a	Cell11:533,JMB149:189,Cell15:1463	AD2 1+ 6039
1,3	late	JMB149:189,PNAS79:1073,PNAS78:7383	AD2 1- 25954
3,5	Ad2 infected cells	Cell22:231,JVIR41:518	XX2 1+ 1853
7	Ad2 infected cells	JVIR41:518	XX2 1+ 287
7	Ad2 infected cells	JVIR41:518	XX2 1+ 873
1,6	early	JVIR30:279,JVIR37:7,JVIR41:449	SV40XX 0- 5233
3,7	early	JMB159:189,JVIR44:175	PAPOA2 0+ 154
1,6	late	JVIR41:449	SV40XX 0- 31
3	late	JVIR44:175	PAPOA2 0+ 22
1,6	late	NAR5:2359,PNAS76:3078,JMB126:813	SV40XX 0+ 325

These numbers are sometimes followed by special characters which indicate that the experiments were performed with RNA synthesized *in vitro* (*), in injected oocytes (°), or in transfected cells (‡). Codes in parentheses refer to promoter evidence from closely related genes. In the column entitled "Expression/Regulation", only the most dominant regulatory features are listed. This information remains fragmentary since many genes are subjected to complex control mechanisms. The literature references given in condensed form refer to the articles on which the assignment of the transcriptional initiation site is based. In some cases, they include reports on transcription studies in experimental test systems or comments on the phylogenetic relationship between the DNA sequence shown here and the gene where the start site has actually been mapped. The rightmost column identifies the nucleotide in the EMBL library sequence which corresponds to position zero in our listing. These references which are used by our programs for automatic DNA sequence retrieval, consist of four elements: Entry name, sequence type (0=circular, 1=linear), strand (+ or -) and position number.

library and organized as a matrix of nucleotides. 2. This matrix is subdivided into overlapping vertical windows (originally termed "cross-sections") which are searched separately for "signal sequences" (oligonucleotides) that are defined in a "signal sequence collection" (*e.g.*, a complete

Table I

Characterization of Constraint Regions by Over-Represented Gapped Trinucleotides.

Eukaryotic Promoters	
TATA-box region (from -35 to -16)	Cap-site region (from -9 to +10)
-30 -25 -20 Occurrence frequency	-5 0 5 Occurrence frequency
< ---TA-A--- > 70.1 % (117/167)	< --C--CA--- > 31.0 % (52/168)
< ---ATA---- > 66.5 % (111/167)	< -CA----C-- > 28.6 % (48/168)
< --A-A-A--- > 63.5 % (106/167)	<---CA-T--- > 28.0 % (47/168)
< --AT--A--- > 63.5 % (106/167)	< --CA--C--- > 27.4 % (46/168)
< ---TAA---- > 61.1 % (102/167)	< --CA---T-- > 25.6 % (43/168)
< --T--A-A-- > 60.5 % (101/167)	<-G----CA-- > 25.6 % (43/168)
< ---A-AA--- > 58.1 % (97/167)	< ---A-CA--- > 24.4 % (41/168)
< --T--AA--- > 56.3 % (94/167)	< ---TCA---- > 24.4 % (41/168)
< --TA---A-- > 55.7 % (93/167)	<---CAG---- > 23.8 % (40/168)
< ---T-TA--- > 55.4 % (92/166)	<---CA-C--- > 23.2 % (39/168)
< ---AT-A--- > 53.3 % (89/167)	< -C------CA- > 22.6 % (38/168)
< --TA--A--- > 53.3 % (89/167)	<---CAC---- > 22.6 % (38/168)
< ---TAT---- > 53.0 % (88/166)	<--CA---C-- > 22.6 % (38/168)
< ---T-AA--- > 51.5 % (86/167)	< -CA-----A- > 22.0 % (37/168)
< ---AAA---- > 50.9 % (85/167)	< G------CA- > 21.4 % (36/168)
< --T-T--A--- > 50.9 % (85/167)	< A------CA- > 20.8 % (35/168)
< --A--AA--- > 49.1 % (82/167)	< --G---CA-- > 20.8 % (35/168)
< --A---AA-- > 47.9 % (80/167)	<-G------CA- > 20.8 % (35/168)
< --T-T-A--- > 47.3 % (79/167)	< --A---CA-- > 20.8 % (35/168)
< -A-A---G-- > 46.1 % (77/167)	<--CA--T--- > 20.8 % (35/168)
< -A-A----G- > 46.1 % (77/167)	
TATAAA	CA-yyy
Prokaryotic Promoters	
-35 region (from -45 to -26)	-10 region (from -19 to 0)
-40 -35 -30 Occurrence frequency	-15 -10 -5 Occurrence frequency
< ---TTG---- > 52.7 % (43/ 81)	< --TA---T-- > 71.4 % (80/112)
< ---T-GA--- > 42.7 % (58/110)	< --A-A-T--- > 56.3 % (63/112)
< ---TT-A--- > 40.0 % (47/110)	< --A--AT--- > 50.9 % (57/112)
< --TT--C--- > 39.1 % (44/110)	< ---TA-A--- > 46.4 % (52/112)
< ---T-TT--- > 38.5 % (43/109)	< --T--A-T-- > 46.4 % (52/112)
< ---TGA---- > 37.3 % (42/110)	< --T---AT--> 44.6 % (50/112)
< --T--AC--- > 35.5 % (41/110)	< --AT--T--- > 42.9 % (48/112)
< TT-------T > 34.6 % (39/110)	< ---AAT---- > 39.3 % (44/112)
< ---TG-C--- > 34.6 % (38/110)	< -G--A---T- > 39.3 % (44/112)
< --T-G-C--- > 34.6 % (38/110)	< ---TAA---- > 39.3 % (44/112)
< --TT---A-- > 34.6 % (38/110)	< --TA--A--- > 38.4 % (43/112)
< --T-T--A-- > 34.6 % (38/110)	< T---A---T- > 38.4 % (43/112)
< -A-----TT- > 33.9 % (38/109)	< TA-------C > 37.5 % (42/112)
< -A----TT-- > 33.9 % (37/109)	< T----A---T > 37.5 % (42/112)
< --T-T-G--- > 33.7 % (37/110)	< ---TA-T--- > 36.6 % (41/112)
< --A---TT-- > 32.1 % (36/109)	< A---T---C- > 35.7 % (40/112)
< --TG--A--- > 31.8 % (35/110)	< --A---TA-- > 35.7 % (40/112)
< ---T-AC--- > 31.8 % (35/110)	< ---T-AT--- > 35.7 % (40/112)
< --T--TG--- > 31.8 % (35/110)	<--A---TA-- > 35.7 % (40/112)
	< ---T-AT--- > 35.7 % (40/112)
t-TTGACa	TATAAT---c

Gapped trinucleotides of total length 10 were searched for in successive overlapping windows of width 14. The signal sequences are listed in decreasing order with respect to their highest local signal frequency as determined in the window that is delineated by angle brackets. Absolute frequencies and local sample size are given in parentheses. The average occurrence frequency of the 2304 signal sequences is approximately 7.5%.

set of trinucleotides). Thereby, the lines where given signal sequences occur are counted in successive windows, a process which yields an integer number called "signal frequency" for each combination of window and signal sequence. 3. The resulting "signal frequency matrix" is processed to final output (constraint profiles, lists of over-represented signals, *etc.*) for localization and characterization of common sequence features. The whole procedure requires specification of a few parameters which also appear in the related methods mentioned above though they have been termed differently. We decided to rename two of them in order to minimize terminological diversity: Thus, the "cross-section length" is now called "window width" in accordance with Waterman et al. (16), and for the "displacement length" we use the term "window shift" as introduced by Schneider et al. (18).

The extensions of signal search analysis include a new search technique described as an option of the ENCODE program of the Delila system tools (18): Usage of "gapped" oligonucleotides (our terminology) as signal sequences. Gapped oligonucleotides are signal sequences in which distinct positions are unspecified. These positions are represented by an additional character (hyphen or **N**) which plays the role of a wildcard. Since statistical analysis of signal search data usually assumes approximately equal occurrence probabilities for all signal sequences, the numbers of both specified and unspecified positions are usually kept constant within signal sequence collections. Moreover, the explicitly specified nucleotides must be centered so that the number of leading N's is either equal to the number of trailing N's or lower by one, in order to avoid multiples of equivalent signal sequences such as **ANANNN**, **NANANN**, *etc.* The gapped dinucleotide collection of total length 6 used for generation of the profile shown on top of Fig. 2 thus consists of all signal sequences of the following types: **NNXXNN**, **NXNXNN**, **NXNNXN**, **XNNNXN**, **XNNNNX**, where **X** can be any of the four bases **A**,**C**,**G**, and **T**. The gapped trinucleotide collections used in our work are defined according to the same principles.

The programs described in (15) allow search for imperfect occurrences of signal sequences. However, in the analyses presented here, it has not been made use of this facility. The parameter "homology limit" is therefore not listed in the legends to the figures and tables. Constraint profiles are shown in a slightly different way as compared to the previous publication (15). Here, we correct the constraint index for the effect the sample size has on the expected variance of signal frequencies. The new index is given by

$$(1) \qquad C_j = \frac{n_j}{(n_j - 1)} \left[\frac{v_j}{m_j(n_j - m_j)} - \frac{1}{n_j} \right]$$

where n_j denotes the sample size, and m_j and v_j the mean and variance of the signal frequencies in the jth window of the DNA sequence matrix. The sample size which varies from window to window is directly reflected by a dashed line on each constraint profile.

The significance of a given signal frequency is calculated as follows:

$$(2) \qquad S_{ij} = \frac{(f_{ij} - m_j)\sqrt{n_j}}{\sqrt{m_j(n_j - m_j)}}$$

where f_{ij} denotes a specific element of a signal frequency matrix. This formula yields only a rough estimate since it does not account for the slight sequence specific variations of signal occurrence probabilities. Its function is to allow comparisons between signal frequencies obtained with different sets of search parameters.

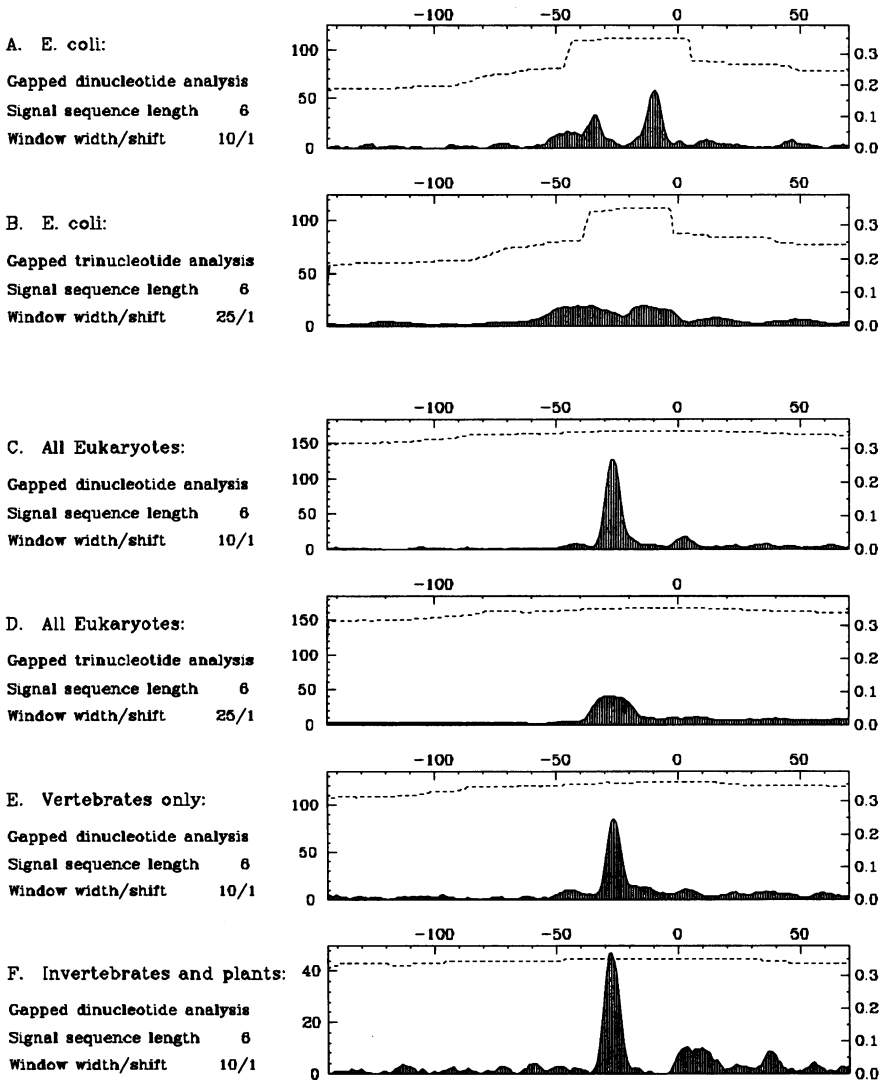

Figure 2. Constraint profiles of *E. coli* and eukaryotic POL II promoters. The curves were calculated as described in the methods section. Dashed lines monitor the local sample size for which the scale is given on the left side. The right-hand labels relate to constraint as defined by equation (1).

RESULTS AND DISCUSSION

We first determined the regions of highest sequence conservation for prokaryotic and eukaryotic promoters by deriving a number of constraint profiles with various signal sequence collections and parameter sets. The general pictures that came up this way were remarkably constant: Two maxima at -35 and -10 for the E. coli system as expected, and one strong peak centered at -28 together with a weak signal near the initiation site for eukaryotes. Two typical profiles are shown in Fig. 2A and 2C. Splitting the eukaryotic promoter set into vertebrate and non-vertebrate sequences revealed only minor differences between these two groups (Fig. 2E and 2F). The cap-site homologies are more pronounced around non-vertebrate transcription start sites. Two additional features can be recognized in the profile that characterizes vertebrate promoters only: A low constraint maximum around -45 and a downstream shoulder of the TATA-peak at -20. These locations coincide with maxima in GC-content (see Fig. 3) and probably reflect only biased base composition.

Constraint analysis allows quantitative comparisons between conserved sequence elements. The profiles in Fig. 2 indicate that the eukaryotic TATA-box is a stronger consensus sequence than the prokaryotic Pribnow-box. In principle, such conclusions cannot be drawn from a comparison of a single pair of constraint profiles, since the relative heights of constraint peaks is much dependent on the signal search parameters specified for the analysis (15). However, in the case of eukaryotic and prokaryotic promoters, we observed that the rank-order of the four dominant constraint maxima (euk. TATA-box, prok. -10, prok. -35, and euk. cap-site) is not affected by changes in parameters (data not shown). We also note that in both systems, sequence similarities are confined to a region extending from approximatly -50 to +10 relative to the initiation site and that total constraint is of a similar magnitude. Integration of the profiles shown in Fig. 2A and 2C within these limits yields values of 2.8 for *E. coli* and 2.7 for eukaryotes. This means that on average two eukaryotic POL II promoters exhibit as many common sequence features as a pair of *E. coli* promoters, and it is a surprising result because the eukaryotic sequence set represents a wide spectrum of organisms, developmental stages and tissues, whereas the *E. coli* sequences are all recognized in an identical biochemical environment. It suggests an extraordinary high conservation of the structure of those parts of the POL II transcription system which are involved in promoter recognition.

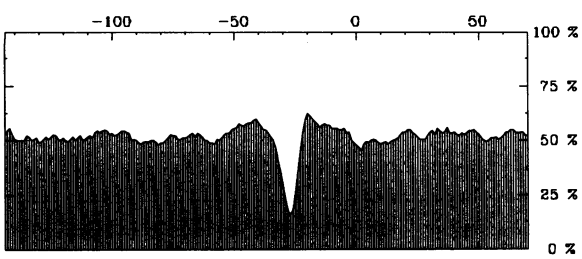

Figure 3. GC-profile of eukaryotic POL II promoters. The base composition was determined in successive overlapping windows of width 5. Similar curves are obtained when the set of sequences is split into vertebrate and non-vertebrate promoters.

In Fig. 2B and 2D we show constraint profiles that have been derived by using wide windows of 25 bases and gapped trinucleotides instead of dinucleotides. Under these conditions it should be possible to detect promoter elements which occur at a more variable distance from the initiation site if they exist in a significant proportion of the analysed sequences. However, for both eukaryotic and prokaryotic promoters these profiles look qualitatively the same as those calculated for narrow windows. The peaks simply become lower and broader. This finding is strong evidence against the existence of any universal consensus sequence upstream from the TATA-box, in other words, there is no -80 region of eukaryotic promoters.

For explicit description of conserved sequence features of eukaryotic and *E. coli* promoters we tabulated the most frequent gapped trinucleotides up to ten base-pairs in total length for the

Table II

Over-represented Upstream Pentanucleotides of Eukaryotic POL II Promoters

Window: -99 - -60			Window: -119 - -60			Window: -129 - -50		
Sample size 157, exp.fr. 3.5 %			Sample size 152, exp.fr. 5.3 %			Sample size 149, exp.fr. 7.2%		
	Frequency	Significance		Frequency	Significance		Frequency	Significance
CCAAT	14.7 %	7.92	CCAAT	17.1 %	6.76	GGGCG	20.1 %	6.52
CAAAA	11.5 %	5.70	CAAAA	15.8 %	6.02	CCAAT	18.8 %	5.87
AGCCA	10.2 %	4.81	GGGCG	15.1 %	5.65	CAAAA	17.5 %	5.21
GGCGG	9.6 %	4.36	AATGA	13.8 %	4.91	AATGA	16.8 %	4.89
GGGCG	9.6 %	4.36	AGAAA	13.8 %	4.91	AGAAA	16.8 %	4.89
AAGGG	8.9 %	3.91	AGCCA	13.2 %	4.54	AGCCA	16.8 %	4.89
AATGA	8.9 %	3.91	CCCCT	13.2 %	4.54	GGGGC	16.8 %	4.89
ACCAA	8.9 %	3.91	CCCGC	13.2 %	4.54	AAAAT	16.1 %	4.56
CTCCA	8.9 %	3.91	TTTCT	13.2 %	4.54	CAGCC	16.1 %	4.56
GCGGG	8.9 %	3.91	AAAAT	12.5 %	4.17	TGTTT	16.1 %	4.56
TGCAT	8.9 %	3.91	CCCCC	12.5 %	4.17	GGAGC	15.4 %	4.23
TGGGG	8.9 %	3.91	GCGGG	12.5 %	4.17	GGCGG	15.4 %	4.23
AAAAC	8.3 %	3.47	AGCAA	11.8 %	3.80	TGTCA	15.4 %	4.23
AAAAT	8.3 %	3.47	ATGAC	11.8 %	3.80	ACCAA	14.8 %	3.91
AGGGA	8.3 %	3.47	GGAGC	11.8 %	3.80	CCCGC	14.8 %	3.91
CCCCC	8.3 %	3.47	GGGGC	11.8 %	3.80	CCTGC	14.8 %	3.91
CCCCT	8.3 %	3.47	TGTTT	11.8 %	3.80	CTCCA	14.8 %	3.91
CCCGC	8.3 %	3.47	TTTTG	11.8 %	3.80	GAAAA	14.8 %	3.91
CCGCC	8.3 %	3.47	ACACA	11.2 %	3.43	GAAAT	14.8 %	3.91
GAAGG	8.3 %	3.47	ACCAA	11.2 %	3.43	GGGGC	14.8 %	3.91
GCGCG	8.3 %	3.47	AGATG	11.2 %	3.43	GTGGG	14.8 %	3.91
GGCAG	8.3 %	3.47	AGGGA	11.2 %	3.43	TGGCG	14.8 %	3.91
GGGGC	8.3 %	3.47	CCTGC	11.2 %	3.43	TTTCT	14.8 %	3.91
			GCAAA	11.2 %	3.43	AAGGG	14.1 %	3.58
			GGGAG	11.2 %	3.43	AGGGA	14.1 %	3.58
			TGGGG	11.2 %	3.43	CCCCC	14.1 %	3.58
						CCCCT	14.1 %	3.58
						CGCCC	14.1 %	3.58
						CGGGG	14.1 %	3.58
						GCAAA	14.1 %	3.58
						GCCTG	14.1 %	3.58
						GCGGG	14.1 %	3.58
						GGGTG	14.1 %	3.58
						TGACA	14.1 %	3.58
						TGGGC	14.1 %	3.58
						TTGCA	14.1 %	3.58

Non-interrupted pentanucleotides were searched for in single windows of width 40, 60, and 80. The significance of the signal frequencies is calculated as described in the methods section.

four major constraint regions shown by the profiles of Fig. 2. Such analysis usually produces clusters of signal sequences which perfectly align to a corresponding consensus sequence (see Table 1). Only in the weakly conserved cap-sequence some positions are occupied by alternative nucleotides. For *E. coli* promoters the consensus sequences reflected by Table 1 are identical to those determined by Hawley and McClure (9) and independently confirmed with computer methods similar to ours by Galas *et al.* (17). The analysis of the eukaryotic -28 region, too, offers no surprise: TATAAA appears as consensus, with the first T being somewhat less important than the other five bases. In the cap-sequence only the dinucleotide CA is well conserved. Otherwise our analysis again suggests a motif which is very similar to previously published consensus sequences for this region (1,2, 19)

Although the constraint profiles of Fig. 2 gave no indication of common sequence features more than 50 bp upstream from the transcription start site, we analysed this region intensively with many types of signal sequence collections and several combinations of search parameters. Special attention was paid to the region where the CAAT-sequence is believed to occur. In general, these analyses did not give very conclusive results. We show in Table 2 the most over-represented non-interrupted pentanucleotides found in three windows of different width. The two oligonucleotides which occupy the top postions are parts of known upstream elements of certain promoters which have been identified by *in vitro* mutagenesis. CCAAT functions in globin genes (20) and GGGCG in the early transcription region of SV40 and in a few other promoters (21). However, as Table 2 demonstrates, the frequencies of these elements are not particularly high as compared to other oligonucleotides which appear in the lists, for instance CAAAA, AATGA, or AGAAA, and their estimated statistical significance is low as compared to the corresponding values obtained for the gapped trinucleotides of Table 1 which characterize constraint regions (the best representatives of the TATA-box and the cap-sequence attain scores of 31.1 and 11.5, respectively). In general, we consider the results shown in Table 2 as supporting the notion that the so called upstream elements and/or enhancers, which are known from experimental studies to play a key role in the expression of eukaryotic genes (for review see 22 and 23), represent a highly polymorphic class of cis-acting genetic elements.

We end our discussion with a few comments on the status of the "CAAT-box". The fact that it cannot be visualized by constraint profiles even with relatively wide windows suggests that the analogy to the -35 region of prokaryotic promoters proposed by Benoist *et al.* (8) is not justified. Moreover, our analysis supports only the functional relevance of the core of the originally proposed consensus sequence GG$^{C}_{T}$CAATCT. It is noteworthy in this context that the pentanucleotide ACCAA which overlaps CCAAT by four nucleotides appears in Table 2 and that mutation of the globin CAAT-box from GCCAAT to ACCAAT results in a threefold increase of promoter activity (24). However, the exact sequence requirements for this upstream element still remain uncertain. It must also be mentioned that an imperfect homology to the sequence CCAAT is likely to be found in an upstream DNA segment of 60bp merely by chance and thus is statistically insignificant. The probability that a given pentanucleotide occurs in a random

sequence of this length with one mismatch allowed is close to 60 % as estimated by equation 1 in (15). It is probable, therefore, that several of the underlined CAAT-boxes in recently published upstream sequences are not real functional analogues of the CCAAT promoter element of globin genes.

References:

1. Corden,J., Wasylyk.B., Buchwalder,A., Sassone-Corsi,P., Kedinger,C., and Chambon,P. (1980). Science **209**,1406-1414.
2. Breathnach,R., and Chambon,P. (1981). Ann. Rev. Biochem. **50**,349-383.
3. Goldberg,M.L. 1979. Ph.D. thesis. Stanford University.
4. Grosschedl,R., and Birnstiel,M.L. (1980). Proc. Natl. Acad. Sci. USA **77**,1432-1436.
5. Benoist,C., and Chambon,P. (1981). Nature **290**,304-310.
6. Yaniv,M. (1984). Biol. Cell **50**,203-216.
7. 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., Proudfoot,N.J. (1980). Cell **21**,653-668.
8. Benoist,C., O'Hare,K., Breathnach,R., Chambon,P. (1980). Nucleic Acids Res. **8**,127-142.
9. Hawley,D.K., and McClure,W.R. (1983). Nucl. Acids Res. **11**,2237-2255.
10. EMBL Nucleotide Sequence Data Library, Release 7 (1985). European Molecular Biology Laboratory, Postfach 10 22 09, D-6900 Heidelberg.
11. Crepin,M. Triadou,P., LeLong,J.-C., and Gros, F. (1981). Eur. J. Bioch. **118**,371-377.
12. Ream,L.W., and Gordon,M.P. (1982). Science **218**,854-859.
13. Langridge,P., and Feix,G. (1983). Cell **34**,1015-1022.
14. Cowie,A., Tyndall,C., and Kamen,R. (1981). Nucl. Acids Res. **9**,6305-6322.
15. Bucher,P., and Bryan,B. (1984). Nucl. Acids Res. **12**,287-305.
16. Waterman,M.S., Galas,D., and Arratia,R. (1984). Bull. Math. Biol. **46**,515-527.
17. Galas,D.J., Eggert,M., and Waterman,M.S. (1985). J. Mol. Biol. **186**,117-128.
18. Schneider,T.D., Stormo,G.D., Yarus,M.A., and Gold,L. (1984). Nucl. Acids Res. **12**,129-140.
19. Busslinger,M., Portmann,R., Irminger,J.-C., and Birnstiel,M.L. (1980). Nucl. Acids Res. **8**,957-977.
20. Dierks,P., van Oyen,A., Cochran,M.D., Dobkin,C., Reiser,J., and Weissmann,C. (1983). Cell **32**,695-706.
21. Dynan,W.S., and Tijan,R. (1983). Cell **35**,79-87.
22. Dynan,S., and Tijan,R. (1985). Nature **316**,774-778.
23. Serfling,E., Jasin,M. and Schaffner W. (1985). Trends Gen. **1**,224-230.
24. Myers,R.M., Tilly,K., and Maniatis,T. (1986). Science **232**,613-618.