Volume 11 Number 13 1983 Nucleic Acids Research

Physical characteristics in eucaryotic promoters

M.Bensimhon, J.Gabarro-Arpa, R.Ehrlich and C.Reiss*

Institut Jacques Monod, CNRS and Université de Paris VII, tour 43, 2 place Jussieu, 75251 Paris, France

Received 25 April 1983; Revised and Accepted 15 June 1983

SUMMARY

For a series of wild type and mutated eucaryotic gene prelude sequences (mainly "promoters" of SV40 early gene (Benoist and Chambon, Nature 290, 304 (1981); Moreau et al., Nuc. Acids Res. 9, 6047 (1982)) and of Herpes Simplex Virus TK gene (McKnight and Kingsbury, Science 217, 316 (1982)), in vivo promoter activity and local stability (denaturability) have been correlated. In agreement with the conclusions drawn in these papers, the correlation points to three major eucaryotic promoter elements and loci: (i) enzyme enabling by an enhancer sequence; SV40 and Moloney Sarcoma Virus enhancers have a striking stability homology; (ii) enzyme activation, occuring 50-70 b.p. upstream the cap site in a high stability domain; the enzyme apparently deactivates exponentially upon moving away to trap site; (iii) enzyme positioning at trap site, 30±5 b.p. upstream the cap site. The trap site contains the TATA box, or, when absent, other low stability domains downstream the activator. The number and occupancy of cap sites may depend on the stability and size of the trap site-cap site couple and its distance from the activator.

INTRODUCTION

A wide body of evidences demonstrates that the organization and regulation of the genetic information flux is at least in part encoded in the nucleic acid template. The DNA or RNA segments known or suspected to be repositories of genetic processing signals (associated for instance with the onset or termination of transcription, translation, replication, ...) are expected to bear characteristic features and were carefully checked therefore. The examples of aminoacid codons and restriction sites, which are written into the template as fixed chemical structures (letter alignements), suggested for some time that the expected signals may appear as specific "keywords". Based on a sample of a few sequences only, this view was first thought to be correct. The sequences of the first procaryotic promoters contained the "keywords" TATAAT and CAATGT, located around 10 and 35 base pairs upstream the transcription initiation site (adresses -10 and -35) (1); the first established sequences preluding eucaryotic genes contained a TATA sequence

also and a CAAT sequence, around adresses -30 and -70 upstream the cap site (2); in procaryotic mRNA, it was found that the AUG triplet was qualified as an initiator codon when associated with a GGAGG sequence located 10 bases upstream (3).

As the size of the sequence bank increased, it became however obvious that these "keywords" were quite often altered. The concept of "consensus sequences" was introduced instead, associating with each letter of the "keyword" a probability of occurence (4,5,6). At present it is evident that even the probabilistic view of consensus sequences is not general, since examples of promoters are known missing the "TATA box" (7) and well-expressed genes exist having in their mRNA no significant element of the GGAGG sequence upstream the initiator codon (5). In contrast, full size keywords are found at places where the alleged, associated genetic function is not detected. Although they certainly play some role, literal homologies are neither sufficient nor necessary to characterize genetic processing signals, which is to say that they are not introduced by any immutable chemical structure.

In many cases, the location and size of the nucleic acid segment bearing the relevant signal are known from mutant studies; these yield precise upper limits within which the signal is deposited: the extensive sequence variation observed therein, even in a given species, indicates that the signal to point must be a physical, collective property, having a priori two characteristics: it can be matched by a variety of sequences and it is spread over a segment of limited size (usually larger than the assumed consensus sequences).

A priori, physical properties of nucleic material, candidates for supporting regulatory signals, could be of diverse nature. Helix handedness or base-per-turn number, the presence of unwound DNA regions at palindromic sites or left-to-right handed helix jonctions (8), lack of nucleoprotein coverage or special higher order folding (9), are regulatory signal candidates. One can also think of dynamic properties particular to nucleic acids (10). Signals could also result from a DNA-specific protein complex; however, the occurence of nucleoprotein complexes at specific sites of the DNA template request also specific recognition signals.

Whatever the physical properties involved, the collective character of the regulatory signals suggest that they will reveal in the thermodynamic properties of the nucleic acid as distinct features.

Furthermore, many -if not all- basic genetic processes operate on linear, single-stranded nucleic acids. Transcription takes place on one DNA strand only, which assumes local DNA unwinding, and so does DNA replication;

translation in the ribosome assembly is likely to proceed on mRNA at least locally devoid of secondary structure, etc... Thus, the state function, associated with the change of state of the nucleic acid (DNA re/unwinding, RNA re/unfolding) could indeed carry characteristic features signaling locally the insert of a given regulatory signal.

These state functions are easily accessible to experimental investigation and can now be computed in great detail from the DNA sequence (11). Recently, we have shown that procaryotic promoters (12) and translation initiation sites (13) distinguish by characteristic patterns in their state functions, suggesting plausible mechanisms for transcription and translation initiation.

The present report summarizes our investigation along this line on putative promoters of eucaryots (viruses mainly) specially SV40 and Herpes Simplex viruses, for which many mutants have been constructed and tested in vivo (14,15,16).

MATERIAL AND METHODS

Nucleic acid helix-coil transition thermodynamics.

Considerable progress has been made in this field when it was recognized that the change of state of nucleic acids (specially DNA) proceeds stepwise. Together with other groups (17-20) we have shown experimentally (see 21 for a review) that in case of thermal unfolding of DNA, each of these steps corresponds to the unwinding of a particular DNA segment or "domain", whose location, size and mean base composition can be accurately derived from the experimental data.

Much effort has been spent to account theoretically for this behaviour from the DNA sequence and its environment (solution conditions). Remarkably, it turns out that two parameters suffice to describe the main features of the process: a parameter "p", related to the DNA "stability" (see below) and an "environment" related parameter "W". If X quantitates some agent able to induce a helix-coil transition of DNA (X may stand for temperature, pH, concentration of denaturing (bio)chemicals, mechanical torque-couple acting on the strands, etc...) then, by definition, $p=(X-X_{AT})/(X_{GC}-X_{AT})$, where X_{AT} and X_{GC} are the values of X for which respectively random poly d(AT) and poly d(GC) change state. W is proportional to V, the free energy associated with an helix-coil boundary; $W=V/(f_{AT}-f_{GC})$, where f_{AT} and f_{GC} are the bound free energies of AT and GC b.p. respectively.

W is very sensitive to the ambient ionic strength cf.(22) . The main contribution to W comes from the change in electrostatic repulsion between the two DNA strands upon helix-coil transition. Calorimetric measurements (23) show that V exceeds by far f_{AT} and f_{GC}, and also the mean energy fluctuations at practical temperatures. This implies that the parts of the DNA molecule, undergoing helix-coil transition for a given value of p, are precisely defined "domains", having characteristic helix free energy matching V. It is for this reason that upon changing state, DNA is partitionned into domains as first recognized by Azbel (24) and later developped in detail by Gabarro and Michel (25) and Michel et al. (26).

From a practical point of view, the main contribution to the partition

function comes from the fundamental state, which allows its numerical calculation by a fast and simple algorithm (25).

The ability of the model to account for the thermodynamic properties of DNA unwinding, is demonstrated by the excellent fit of the experimental melting profiles with those calculated from the sequence, according to the model. This remarkable agreement is observed within the W range experimentally accessible at present (W from 5 to 10, corresponding roughly to ionic strength from 10^{-1} to 10^{-3} M of monovalent ions) and for all DNA species studied so far, both eucaryotic and procaryotic (27) including DNAs with extreme mean base composition constrains, like yeast mitochondrial DNA (26). The same agreement was observed by Benight et al. (28).

A suggestive picture of the change of state behaviour of DNA is given by a three-dimensional plot, termed "DNA state surface", derived from the DNA sequence by the algorithm of Gabarro and Michel (25) : DNA sequence (S) vs. stability (p) vs. environment (W). An example is given in figure 1, on which are also visualized the "stability profile" (S vs. p at constant W), the "environment map" (projection of the DNA state surface on the (W,S) plane) and the DNA phase diagram (DNA state surface projection on the (p,W) plane). The DNA melting domains are displayed on the stability profile as bars of constant p, the stability at which individual domains change state. The environmental map shows how the DNA molecule subdivides into smaller domains as W increases.

In what follows, we will examine the stability profile of DNA in environments at W=4. Although we have so far no experimental proof that our model remains valid below W=5, we postulate this to be the case. Special experiments to test this hypothesis are currently being designed.

We have computed the DNA stability surfaces of putative promoter sequences gathered from references (14,15,16,29). Computations were carried out on a UNIVAC 1110 computer at the University of Paris XI.

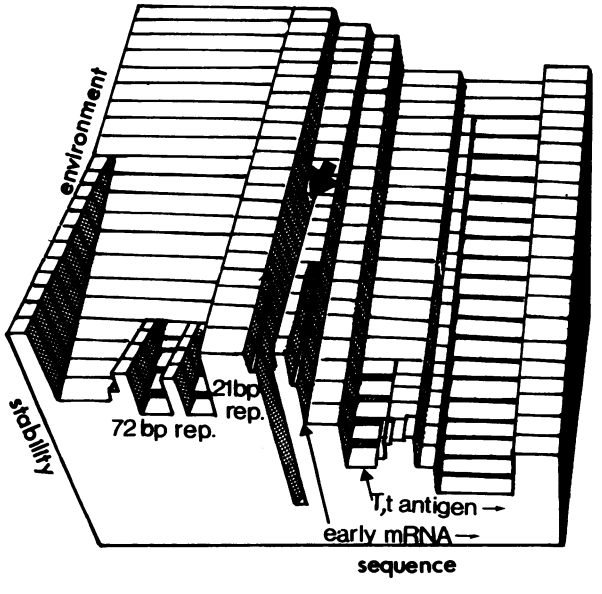

Figure 1.

Three-dimensional stability diagram for the sequence of a fragment of SV40 starting 24 bp. upstream HpaII site and terminating 781 bp. downstream (see text).

RESULTS.

For reader's convenience, selections of results given in (14), (15), (16) and (29) are briefly reproduced below, at pertinent places in the comments of stability profiles. This ad-hoc selection is neither exhaustive, nor does it infere a critical appraisal on the results not mentioned.

I-SV40 AND RECOMBINANTS DELETED IN VARIOUS PLACES OF THE PUTATIVE EARLY GENE PROMOTER REGION (fig. 2).

1) Wild type SV40.

Figure 2 displays the stability profile of a SV40 segment 727 base pairs long (see legend to fig. 2) including the early gene promoter region. When compared to the stability profile of the complete SV40 genome (data not shown), this part of the profile is remarkable in that it bears by far both the most stable and the most unstable domains of the whole molecule. The "TATA box" is located in a domain of absolute stability minimum (p=.25) which is not surprising since 23 of 27 b.p. at the "TATA box "site are AT pairs. Its immediate neighbour domains are in contrast the most stable domains of the molecule (p=.7). The caped site is precisely located at the distal boundary of the stable domain downstream the "TATA box".

Another striking feature is the presence of two 20 b.p. long stable domains R^{I}_{20} and R^{II}_{20}, each being part of the 72 b.p. perfect repeats located within the early-late intergenic region of SV40 and known to play an important role in the early gene expression (see below). Notice that the R^{I}_{20} and R^{II}_{20} domains come out into view only below W=4.5 (see fig. 1) and remain of the same size down to W=2.0 (data not shown).

We notice incidently that the initiation codon of t and T antigens is in a local stability dip (p=.35) a feature found for all (except one) procaryotic and eucaryotic initiators examined so far. We have proposed that this is the key element qualifying an ATG to be an initiator codon (13).

2) Mutants deleted between the "TATA box" and the caped site (14).

Here, the stable domain just downstream the "TATA box" is progressively reduced in size and lowered in stability. The main cap sites of these mutants remain located 30 ± 5 b.p. downstream the box, except for HS4, for which several secondary cap sites clearly show up downstream (site 1 to 11 in fig. 2, see legend to fig. 3c in (14)). T antigen production of these mutants is only slightly reduced in comparison with wild type expression.

3) Mutants deleted between the caped site and upstream the "TATA box" (fig. 3) (14).

From the point of view of T antigen expression, two classes can be

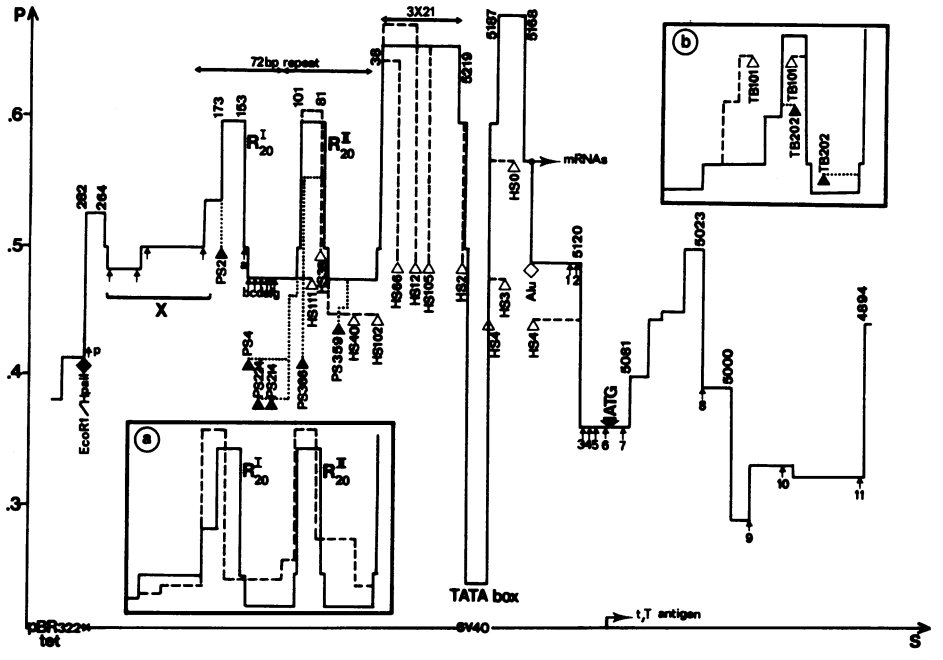

Figure 2.

Solid line: stability profile (stability p vs. sequence for W=4.0, see METHODS) of 750 bp. fragment of pSV1, including 23 bp. from pBR322 (TET side of EcoR1 site) and 727 bp. from SV40 (early/late intergenic region from HpaII site and beginning of early gene). Sequence and numbering of main domain boundaries (see METHODS) as in (14). The perfect 72 bp. repeat, and the three nearly identical 21 bp. repeat (3X21) are indicated.

◇ . Alu site at 5162, which is the downstream deletion point common to all HS mutants.

◆ . EcoR1/HpaII site, which is the upstream deletion point common to all PS mutants.

△ . upstream limits of HS mutant deletions.

▲ . donwstream limits of PS mutants deletions (mutant identification indicated below these signs).

Dashed lines: part of the stability profiles of HS mutants.

Dotted lines: part of the stability profile of PS mutants.

The stability profile of any HS mutant (except HS4, see fig. 2) is composed of the solid line part downstream◇, the dashed line starting at the point where the dashed line merges into the solid one. PS mutant stability profiles are constructed in the same way.

Arrowheads: caped-sites(taken from (14)); horizontal: pSV1 mRNA start site; vertical, P and a to g: common to mutants HS38 and HS111; vertical, in bracket X: for HS111 only; vertical, 1 to 11: common to TATA minus HS mutants up to and including HS38.

Insert a: full line, stability profile of pSV1 in the 72 bp. repeat

Fig. 2 (continue) region (as above); dashed line: part of the stability profile of SV r^{MSV} determined by the inserted MSV 72 bp. repeat (see (29)).

Insert b: full line, stability profile of recombinant TBO in the region of the remaining copy of the 72 bp. repeat; dashed line, stability profile of mutant TB101, resulting from TBO by deletion between the two Δ ; dotted line, stability profile of mutant TB202, resulting from TBO by deletion between the two ▲ .

distinguished in these mutants. Those not extending into R^{I}_{20} (HS2 to HS40) are characterized by a reduction of size and stability decrease in the stable domain immediately upstream the "TATA box" (which contains three 21 b.p. repeats, dubbed 3X21). Their in vivo T antigen production progressively reduces as the deletion becomes more extended; for instance, HS40, the most deleted member of the class, has only few percent of the wild type (w.t.) T antigen production. Inspection of the transcriptional activity map shows that this lowered phenotypic activity is partly due to a progressively increasing amount of cap sites donwstream the T antigen initiator. For instance, for mutant HS66, less than 60% of the mRNA include the initiator, with antigen T level of 15%.

The second class of these mutants extend into and beyond R^{II}_{20}. The stability profile of HS38 (deletion extending 3 bases into R^{II}_{20}) still shows a remnant of the R^{II}_{20} domain (stability reduced from p=.6 to p=.55) but the R^{II}_{20} domain is completely erased in mutant HS111, which extends 11 further into R^{II}_{20}, as well as for more deleted HS mutants (not shown). The T antigen expression of HS38 compares to that of HS66, whereas HS111 and the following produce no T antigen in short-term experiments (see (14)). These two classes of mutants share in common the series of 11 discreted caped sites (1 to 11, fig. 2). Remarkably, the loci of these sites (fig. 2) are mostly within stability dips, p<.5. However, the occupancy of these sites is neither homogeneous for a given mutant (occupancy distribution is approximately bell-shaped) nor identical for all mutants (occupancy distribution maxima remains roughly 30 b.p. downstream the deletion point). We notice that for HS38, 45% of the transcripts start in X region (5 bands, fig. 2 and see fig. 6 of (14)) that is a region upstream R^{I}_{20} and downstream the stable domain 282-264 , flanking the EcoR1 site on the Tet side. HS38 and HS111 share in common 7 mRNA bands between R^{I}_{20} and R^{II}_{20}.

4) PS mutants (14).

Mutants PS2 and PS4 delete respectively part and all of the R^{I}_{20} domain; their transcriptional behaviour compares to that of the wild type. PS224 and PS214, which bear deletions further downstream but not into R^{II}_{20}, still

support gene expression, although with reduced yield (10%). Mutants PS366 and PS359, which deletes part or all of R^{II}_{20} do not support gene expression in short-term experiments.

5) TB mutants (15).

Mutant TB0 has lost the equivalent of one complete 72 b.p. repeat. Transfection of this mutant is 85% of that of the w.t. Further alterations of this mutant are introduced by deletion starting approximately in the middle of the remaining R_{20} sequence and extending both upstream (32 b.p., mutant TB101) and downstream (21 and 22 b.p., mutants TB208 and TB202. In these mutants, gene expression is virtually abolished (4% from w.t. in TB202 and TB208, <1% in TB101). The stability profiles of TB mutants are shown in fig. 2, (insert b). The TB0 profile is identical to that of the wild type except that R^{II}_{20} is missing. The remaining R^{I}_{20} is altered (shrinkage and stability lowering) in TB202, TB208 and TB101, the later mutant showing a remarkable stability increase in the domain upstream R^{I}_{20} due to predominant AT deletion there.

II-RECOMBINANT SV-r^{MSV} (29).

This recombinant results from the replacement in SV40 of the 72 b.p. repeat by a 72 b.p. repeat derived from the long terminal repeat (LTR) of Moloney murine Sarcoma virus (MSV) DNA. The SV40 and MSV 72 b.p. repeats bear no detectable sequence homology. SV-r^{MSV} induces T-antigen synthesis in monkey cells as does wild type SV40. Fig. 2 (insert a) shows the stability profile of SV-r^{MSV} superimposed to that of pSV1. Notice the close resemblance of both profiles in the 72 b.p. repeat regions.

III-SUBSTITUTION MUTANTS UPSTREAM THE CAP SITE OF HERPEX SIMPLEX VIRUS (HSV) THYMIDINE KINASE (TK) GENE (16).

This series of mutants result from 5-10 nucleotide substitutions clustered within spots located from -119 to + 15 b.p., respectively up and downstream the putative cap site of the TK gene (these mutants, produced by a "linker scanning" (LS) method, are nomenclatured by two numbers standing for the upper and lower limit of the substitution cluster). The stability profile of wild type HSV TK and of strong up or down LS mutants are shown in fig. 3. Some substitutions are AT to TA or GC to CG changes, which are not taken into account in our present calculations of stability profiles.

Prominent features of HSV TK wild type stability profile are (i) the stable domain (p=.70) extending from -105 to -93; (ii) the stable domain (p=.8) extending from -57 to -40; and (iii), the unstable (p=.35) TATA domain extending from -27 to -20. Measurements of the in vivo gene expression of LS

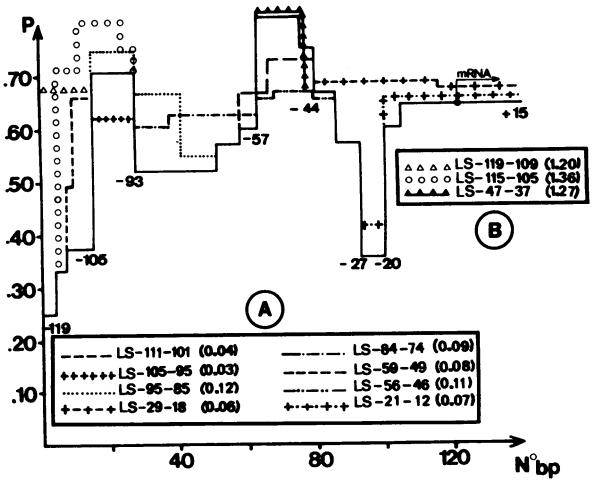

Figure 3.

Stability profile of HSVTK prelude sequence (solid line) and various substitution mutants from 120 bp. upstream the caped site to 15 bp. downstream (16). Numbers refer to domain boundary adresses with respect to caped site. Insert A: down mutations; insert B: up mutations. Numbers in parenthesis behind mutants name are the corresponding in vivo expression with respect to non-mutated HSVTK.

mutants relative to the wild type, show strong alteration for mutants substituted between -111 to -74, -59 to -46 and -29 to -12.

DISCUSSION.

In what follows, our goal is mainly to correlate in vivo transcription data presented at various places -specially in (14), (15) and (16)- with the stability analysis of the corresponding promoters, in order to try to shed another light on the facts and perhaps to bear out another look at the eucaryotic promoter problem.

ROLE OF THE "TATA BOX".

For SV40 early genes, deletion of the "TATA box" does not seem to affect quantitative early gene expression (see mutant HS2), nor does its presence warrant transformation (see mutant PS366). Benoist and Chambon's in vivo experiments, "clearly show that the TATA box has a key role in fixing the start site within a narrow area" (14). Our analysis allows to correlate this role with local DNA stability and to sketch a plausible mechanism by which the TATA area (or low stability substitute) brings the transcriptional complex to fall into step, forcing transcription to start at a given site.

For wild type, pSV1 and for 4 out of the 5 mutants retaining the "TATA box" (P1A, AS, HS0, HS3, see(14)) transcription start almost exclusively within a small area, 30±5 bases downstream the box. The fifth, HS4, which has lost the stable domain downstream the "TATA box", starts to transcribe at the secondary cap sites 1 to 11. This suggests that the deleted GC-rich domain

contributes to the determination of a single start site.

In addition to other effects (see below), the stable 3X21 domain upstream the box may have a similar action, as its progressive deletion (HS2 to HS66) brings upstream the maximum of the start site distribution.

The particular base distribution (and the corresponding stability profile) of the SV40 early gene "TATA box" and stable flanking domains, seems thus to act as a trap, which brings all enzymes to start transcription at the same place. The sequences of many others TATA box area have comparable base distribution constrains; striking examples are the TATA area of adenovirus-2 major late gene (7), 5'-GGGGGGCTATAAAAGGGGGTGGGGGCGCG-3', chicken conalbumine (45), 5'-GCCAGGGCTGCTCCTCTATAAAAGGGG-3', herpes simplex virus TK (16), 5'-CCGAGGTCCACTTCGCATATTAATGACGCGTGTGGCC-3'. Since these genes exhibit single cap sites, their TATA area could also fit the postulated trapping action.

Conversely, the TATA minus mutants, or natural genes lacking TATA like sequences (adenovirus-2, 72K DNA binding protein gene for instance (7)) show multiple cap sites which may result from the absence of a dominant TATA trap.

How does the trap work? The main virtue of the "TATA box" could be its AT-richness, that is its easy denaturability to form a "bubble". In contrast, the stable GC-rich flanking domains bear strong double helices, which provide abrupt boundaries to limit the "bubble" extension. If some element of the transcriptional machinery has an affinity for unwound DNA (as is the case for E.Coli RNA polymerase (30)) the precisely positioned "bubble" could trap the enzyme on the template, thus defining the cap site. The extreme stability constraints observed around the trap could be requested for the processing of this promoter in a wide range of environmental conditions and/or topological states of DNA. This would resemble a situation already encountered in procaryotic promoters (R. Ehrlich et al. manuscript in preparation).

The postulatedtrapping action of the TATA region through the "bubble" effect is consistent with the fact that for the five deletion mutants retaining the "TATA box" (P1A, AS, HSO, HS3, HS4) "initiation always occurs 27-34 base pairs downstream the first T of the TATTTAT sequence" (14), just as for pSV1.It is as if the trapped enzyme had its active site protruding so as to be in contact with the template roughly 3 helix turns downstream.

Transcription of TATA-minus HS mutants provides further support for this scheme. The discrete cap sites 1 to 11 found in mutants HS4 to HS111, the additional sites a to g found in both HS38 and HS111 and the five bands located in zone X for HS38 (fig. 2), share a remarkable property: each is preceded 30 ± 5 bases upstream by an AT-rich box, which could act as a trap

(see table I). For these sites, it would be of interest to correlate the efficiencies of the individual cap sites with the stability and/or extension of the domain covering the related trap substitute. Another remarkable feature is that all 23 cap sites -except numbers 2 and 3 of the 1 to 11 series- are precisely positioned in distinct stability wells, in contrast to the early gene mRNA cap sites of TATA box plus species. Identical conclusion are reached for other TATA-minus "promoters" in the literature (ref. 31 to 34; see table I).

In our hypothetical scheme, this observation could be interpreted as follows: high positioning efficiency of the enzyme by the genuine TATA, allows transcription to start even in a stable domain; less efficient steric positioning of the enzyme by trap substitutes, due to less adequate size or stability, requests that the active site of the enzyme be in contact with a low stability domain, easily denaturable. A similar mechanism has been postulated for E.Coli RNA polymerase melting into procaryotic promoters, in particular to account for the behaviour of the Lac w.t. promoter in the presence of CRP, as compared to that of the mutated Lac "up" promoter, UV5 (12). According to the proposed mechanism (R.Ehrlich et al., manuscript in preparation) CRP is requested on the w.t. promoter to achieve a very precise steric positioning of RNA polymerase, in order to allow its σ subunit to be in contact and to destabilise the GC pairs located at precise adresses just upstream the start site. The result, an unwound 11 b.p. loop including the transcription start, can be obtained also in the UV5 promoter, by σ denaturing only two GC b.p., chosen at will in a set of six; this means less stringent steric positioning of the enzyme, hence the CRP independence and high efficiency of the UV5 promoter.

The assumed trapping action of the "TATA" box, allows to rationalize the behaviour of mutants LS-29-18 of HSVTK (16). Although only modified in its TATA box, this is a strong down mutant, infering that in this case, the TATA box is an essential promoter element. However, close inspection of the DNA sequence or stability profile between the "TATA" box and over 200 b.p. downstream, that is well beyond the TK initiation codon, reveals no trap substitute nor any trap-start couple (the AT richest cluster, a TTAA sequence in a GC-rich environment, 30 b.p. downstream the cap site, does not particularize in the stability profile (not shown)). Thus, in this mutant, the anticipated trapping action takes place very poorly, at least upstream the initiator, explaining consistently the very low in vivo gene expression. A similar element could at least contribute to the "down" effect of LS-21-12;

although the substitutions are outside the TATA sequence, they teleact into the TATA domain (-20, -27), see fig.3.

The postulated TATA trap mechanism is consistent with the experimental results available, but remains a hypothesis. It is also plausible that the TATA area participates to promoter action through other properties and that the precise position and nature (purine-pyrimidine) of some bases of the box are of importance. However, in vivo, this area is obviously not the repository of the signal pointing the SV40 promoter to the RNA polymerase nor does it activate the enzyme-template complex: full level transcription activity is observed in the TATA-minus mutant HS2, whereas PS366, which retains the "TATA" box, does not support gene expression.

A CRUCIAL GC-RICH DOMAIN UPSTREAM THE "TATA" BOX.

The transcription behaviour of mutants HS2 to HS 102 demonstrates that the 3X21 domain behaves as a major transcription efficiency control for SV40 early gene expression. Size reduction of this domain with its stability unchanged ($p=.67\pm.02$) shows a concomitant transcription efficiency lowering. HS2, keeping a full size 3X21 domain, has almost 100% transcription efficiency (80% gene expression, 20% post-T antigen initiator cap sites (PICS)). Size reduction by 1/3 (HS105, deletion of the most downstream repeat) lowers transcription to 65-70% (55% gene expression, 30% PICS estimated from interpolation of HS2 and HS12 data). Deletions of 1/2 (HS12), 5/6 (HS66), or all (HS102) of the 3X21 domain yield comparable gene expression ($15\pm1\%$) and transcription efficiencies around 20% (PICS respectively 35%, 42%, and 50% (the later extrapolated from the two former data)).

Comparison of the mRNA yields of HS2 (100%), HS105 (65%) and HS12 (20%) focusses special attention on the sequences between the 5'deletion points of these mutants, 5'-AGTTCCGCCCATTCT-3' and 5'-TAACTCCGCCC-3' located respectively 30 and 40 b.p. upstream the "TATA" box. The common sequence -TCCGCCC-, may prove an important element in that its deletion reduces transcription by almost 50%.

We have looked for the sequences located 20-40 b.p. upstream the "TATA" box in various eucaryotic (viral and cellular) "promoters" (table II). No obvious sequence homology is apparent but, strikingly, we found in the stability profiles, almost systematically, domains of very high stability composed of clustered GC pairs, extending over five bases at least, often exceeding 8 and sometimes intermingled with just a single AT pair.

The transcriptional behaviour in vivo of mutants in the "A" domain of various promoters mentioned in table II have been reported. For instance,

TABLE I.

start site number (14)	mRNA start	TATA substitute domain	Adress of first base of TATA substitute domain
1	TTTTTT	TATT	-27
2	GGCCTA	AGAAGTA	-28
3	GGCCTA	AAGTA	-28
4	TTTT	AGTAGT	-32
5	AAAAA	TTTTTT	-27
6	TTTGCAAA	TTTTTT	-32
7	TTTTAAA	AAAAA	-28
8	AATATT	TAAT	-34
9	AATGCAA	AATATT	-33
10	ATAAAG	ATATTTAAAAAAAT	-32
11	TAAAATA	TAAAATGAATA	-34
species	**mRNA start**	**TATA substitute domain**	**Adress of first base of TATA substitute domain**
sea urchin H2A mutant .B (31)	ATTCAA	TTCACAA	-25
	TCACAAT	ATTCAA	-29
	ACAATT	AACCATT	-29
rabbit .globin .-100 (32)	TTTTGACACAA	AAAACAG	-33
.-34-4 (32)	ATTACATA	TTACACTT	-26
chicken lyzozyme (33)	AATCAAAA	AAGA	-31
	TATACTCAA	AGAACAGA	-37
SV40 late 1 (34)	ACTTT	TAATT	-26
SV40 late 2 (34)	ACTTT	ACACACATT	-24
SV40 late 3 (34)	TAACCAAGTT	TTATTT	-26

TABLE II.

Species/Gene	ref.	"A" domain sequence	distance* trap domain-(TATA box)
Mouse/β^{min}Globin	(36)	GTGGGCAGGAGCC	-20
Mouse/β^{maj}Globin	(36)	GAGGGCAGGAGCC	-25
Mouse/αGlobin	(37)	GGGCGTGTCCACCCTGCCTGG	-30
Goat/Globin β^A, β^C, γ	(38)	GGGAGGGCAGGAGGCAGGGC-TGGGC	-17
Hen/Ovalbumin	(39)	CCTGTGGGTGGGTC	-24
Rabbit/.Globin	(40)	CGGGGTAGGG	-30
Human Chromosomal/ Interferon IFN-α1	(41)	GTGGCCCAG	-38
Adenovirus2/ Gene of protein IX	(42)	TGTGTGGCGTGGC	-31
Adenovirus2/16.3 leader	(43)	GTGACCGGGTC	-26
Sea Urchin/Histone H2A, h22	(44)	GCCCTCGCTGACCGG	-34
" , h19	(44)	GCCCCCCGACGG	-33
Sea Urchin/Histone H2B, h22	(44)	CGCATCGCAGC	-25
" , h19	(44)	CGGACCGCAGC	-32
Sea Urchin/Histone H3, h22	(44)	CGGTGACGGCGGTGCG	-28
" , h19	(44)	GGGCGACGGCACCAGG	-27
Sea Urchin/Histone H4, h22	(44)	CGAGCCTCGTCATCC	-29
" , h19	(44)	GCCGAATCTCGTCACC	-30
Hen/Conalbumin	(45)	CACAGCCAGGGCTGCTCC	-17
Moloney Sarcoma/5'-LTR	(29)	GCGTTCTGTTCGCGCGC	-30
Drosophila/hsp70	(46)	GCGCGGCCTGCG	-37
S. Purpuratus/Histone H1	(47)	CGGACGACCCGGGACTGTCTC-CTCCC	-35
Adenovirus7/28K protein gene	(48)	GTCCTGTG	-39
Polyoma/early	(49)	CCGCAGCTGGCCGTGCG	-26
Human/α_1 Globin	(50)	GCGCCGCCCGGCCGGGCGTCGCCCC	-35
Adenovirus2/major late	(7)	CCGGGTGTTCC	-20

* DISTANCE BETWEEN "A" DOMAIN (UNDERSCORED BASE) TO MIDDLE OF TRAP DOMAIN.

strong down mutations are observed upon lowering the stability of the "A" domains preluding the drosophila heat shock protein 70 gene (mutants pHT6, pHT7, AS'-44 (51)) or the rabbit B-globin gene (58 (deleting upstream from the middle of the "A" domain), 34 (deleting the "A" domain in full and sequence downstream), (52)). However, these observations deal with deletion mutations which often cover more than the "A" domain, rendering meaningful correlation of the observed effect and the sole "A" domain questionable.

A more convincing argument is derived from the behaviour of the three mutants substituted in the "A" domain of HSVTK gene (16). One (LS-47-37) expresses the TK gene at 127% of w.t., whereas LS-59-49 and LS-56-46 transcribe at 8% and 11% of w.t. respectively. The stability profiles show that the "A" domain is slightly more stable and extended in LS-47-37 (127% of w.t., "up" promoter) but is of lowered stability and reduced size for the two down mutants (fig. 3).

Taken together, the transcriptional behaviour in vivo of the HS mutants of SV40 and those just described, allows to grasp the elements of the "A" domain which seem important for the role it obviously plays in transcription control. A priori, the data suggests the following elements as potentially critical: "A" domain sequence; its base composition and related physical parameters (stability, helix conformation); its size; its relative position in the "promoter". The sequence per se (i.e. the simple letter alignement) seems not enough to secure the function ascribed to the A domain. There is no sequence homology among the "A" domains in table II, in particular among those known to be active in controlling gene expression in vivo. Even more convincingly, in HS12, comparable transcription levels are exhibited by HS12, which has still one copy of the 21 b.p. repeat, and HS44 or HS38.

The three mutants substituted in the "A" domain of HSVTK gene, as well as substitution mutants pRE4 and pRE7 of SV40 (15), which have "A" domains of reduced size (not shown) due to GC to AT substitutions there, show that the high GC content of this domain is an important element, a conclusion also supported by the "A" domains in the eucaryotic promoters (see table II). The precise property conferred by the high GC content on the transcription control is unknown and awaits physical studies of the "A" domains. The concomitant decrease of transcription efficiency and "A" domain size in both deletion mutants (HS2 to HS102) and substitutions mutants (pR4, pR7, and LS-59-49, LS-56-46) infers a direct link between the anticipated "A" domain control function and its size, i.e. the lenght of the stable domain (p=.6 to .7) located about 50-70 b.p. upstream the cap site. The typical size of the "A"

domain is difficult to determine, as the length of those enlisted in table II varies within large limits (average 15 b.p.). LS mutants of HSVTK around the "A" domain clearly delimitate a segment of about 20 b.p. harbouring the function we ascribe to an "A" domain, outside which mutations do not have great effect on gene expression. This size is about 1/3 that of the SV40 early gene "A" domain (64 b.p.). However, since no figures of absolute transcription efficiencies of both genes are available, and also since the stabilities of their "A" domains differ significantly, correlation of their sizes is questionable. Finally, since the copy left in HS12 is almost inoperant (HS38 still supports 10% transcription), and since the two copies present in HS105 only promote transcription to 65% of the level observed in HS2 (which has the three copies), the two donwstream copies, and specially the one flanking the TATA box, seem dominant. This infers that their position, relative to some upstream element, could be critical for transcription control. No data exist as to the location or nature of this element (see conjecture below).

An indication as to the precise function of the "A" domain is given again by the SV40 mutants. That it is not sufficient by itself to determine transcription is demonstrated by TB101, PS366, PS359 for instance, which have afull size 3X21 domain and "TATA" box but express antigen beyond 1% of w.t. only. Two series of experiments allow to state that the "A" domain in SV40 may control transcription level in behaving as an activator. First this role would be consistent with the data on the location of the mRNA start sites and the distribution of the amount of mRNA for each of these sites in mutants HS2, HS12, HS66. As long as at least one of the three repeats is present (HS2, HS12) the start sites are in or downstream of what remains of the 3X21 domain. Already in HS66, a few percent of the mRNAs start just upstream the 3X21 vestige. This means that here, and a fortioti upstream in HS38 and HS111, activation obviously takes place upstream (we suspect (see below) that the elements responsible for this are the two stable R_{20} domains and the stable domain of 17 b.p. on the early side of the EcoR1 site of SV40).

The distribution of the amount of mRNA starting at the various cap sites in HS2, HS12, HS66 is also remarkable, as it seems always centered roughly 30-40 b.p. downstream the remnant of the 3X21 domain, a feature which, as we shall see now, is also consistent with an activator role of the latter.

The second series of experiments strongly suggesting an activator role for the "A" domain, come from the transcriptional behaviour of mutants in the SV40 early promoter region (15), in which pBR322 fragments of increasing length have been inserted between the 3X21 domain and the "TATA" box (mutants

pRE7, pMMB28, 14, 1, 24 and mutants pRE7BS, pMMBS8BS and pMMB1BS).

Correlation between the size x of the insert and the measured transfection level E, shows that a single exponential, $E=E_0\exp(-kx)$ with $k=1/110$, fits the data quite well. This simple relationship suggests that once activated on the 3X21 domain, the activity of the enzymatic complex decays steadily (e^{-1} drop upon shift over 110 b.p.) according to Poisson statistics, as it moves towards the trap sites. This may be the reason why the "A" domain always preceeds the "TATA" box by 20-40 b.p. or why the maximum amount of mRNA is produced roughly 30-40 b.p. downstream the 3X21 remnant in HS2, HS12 and HS66 (these 20-40 b.p. distance may be requested because of the topology of the enzyme or could be related to its activation kinetics). In HS38, the complex mRNA distribution pattern reveals three maxima (14): one in the "x" domain upstream R^I_{20}, one around R^{II}_{20} and one downstream the antigen initiator, the later being absent in HS111, which lacks R^{II}_{20}. The position of these maxima is consistent with enzymatic activation on the stable domain (17 b.p.) just on the early gene side of the EcoR1 site of SV40 (sequence GGCGCAGCACCATGG) which could act as an "A" domain substitute for transcripts starting at P and the seven bands between R^I_{20} and R^{II}_{20} (see legend to fig.5 in ref. (14)); activation could mimic an "A" domain for transcripts starting around R^{II}_{20} ("A" domain at R^I_{20}) and within the 11 start sites ("A" domain at R^{II}_{20}) respectively.

It is likely that other GC-rich sequences in pBR322 could also act as "A" domain substitutes and could be responsible for enzymatic activation in a number of mutants lacking the 3X21 domain (i.e. the pEMP series (15)).

ENTRY OF THE TRANSCRIPTIONAL MACHINERY INTO THE TEMPLATE.

As already mentionned, the presence of the "A" domain is not sufficient for transcription to take place. Mutants HS and TB of SV40 (14, 15) show that the transcriptional determinant enabling transcription must be located upstream the 3X21 domain. Convincing arguments (that need not be rediscussed here) have been given by Moreau et al. (15), that possibly in SV40 "the 72 b.p. sequence corresponds to a particularly efficient bi-directional entry site for a component of the transcriptional machinery". A single copy of the 72 b.p. repeat is sufficient to secure this function (TBO, (15)). Functional replacement of the two 72 b.p. repeats of SV40 by those derived from the LTR of MSV (29), shows that the sequence responsible for entry is not unique, nor does any sequence homology play a role.

We notice, however, a striking resemblance of the stability profiles of the 72 b.p. repeats of SV40 and MSV, each exhibiting a stable ($p<.6\pm.02$)

domain of about 20 b.p., rising over an area of medium stability and centered about 100 b.p. upstream the "TATA" box. In HSVTK, mutants substituted in the region -85 to -50 b.p. upstream the "TATA" box, express the TK gene only to a few precent of w.t. The stability profile of this region also shows a stable (p=.7) domain of 12 b.p., centered about 70 b.p. upstream the "TATA" box, and rising over an area of medium stability. By analogy with the 72 b.p. segments of SV40 and MSV, this region could also be requested to enable transcription of the TK gene. Many other prelude sequences of eucaryotic genes contain, roughly 70 to 150 b.p. upstream the "TATA" box, GC-rich sequences (manuscript in preparation). The position of these putative enabling domains with respect to the "TATA" box should not be considered as critical, since it has been shown in SV40 that the entry function can act at kilobase distance (TBB101 for instance in (15)). Furthermore, in contrast to the "A" domains, it does not seem that the stability of the enabling domain "E" is the (only) important feature of its assumed "entry" function, at least for HSVTK (16): down mutations in the supposed enabling domain of HSVTK either do not change its stability (LS-84-74, LS-111-101), lower it (LS-105-95) or increase it (LS-95-85). More subtil properties must be involved, but the absence of literal homology andspecially the bi-directional ability of the 72 b.p sequence in SV40 points obiously to physical, collective properties of the sequence for the assumed entry function.

Finally, although no clear experimental evidence is available, we suspect some extent of feedback between the "E" and "A" domain, in SV40 at least. As noticed earlier, it is likely that the relative position of the downstream copy of the repeat in the 3X21 domain is critical for an efficient activation. This element may be part of the 72 bp segment. In TB202 and 208 for instance (4% gene expression), the entry function is present (to some extent at least) but expression (i.e. activation) is impaired, due either to the deleted sequence or to the reduced distance between the 3X21 domain and some relevant element upstream. This is perhaps the remnant of R_{20}; in this case, if a TBB202 or 208 exist, i.e. plasmids obtained by blunt ligation of pHS102 box fragment A into the PvuI site of pBR322 part of TB202 or 208 (see (15), it should express as TB202 or 208, in contrast to TBB101 which expresses 50% of w.t., whereas TB101 does not express at all. This hypothetical tandem action of the two domains "A" and "E" resembles the action postulated in procaryotes between the two stable domains flanking on both sides the recognition site (13); these provide attachement sites to the RNAP

prior to its melting into the template downstream to the Pribnow box (see (13) and R. Ehrlich et al. manuscript in preparation).

CONCLUSION.

Promoter recognition and initiation of transcription by the eucaryotic enzyme machinery do obviously not depend on any fixed chemical structure (keywords or consensus sequence). They are most probably signaled by collective properties of base pair sets of finite size, which can be shared by genome segments of different sequences. We believe that these collective signals are the ultimate targets sensed, recognized and used by the transcriptional machinery.

DNA stability (denaturability) has been assayed here as a relevant collective signal candidate. Its correlation with in vivo transcription data bears out a hypothetical scheme for the physical structure of eucaryotic promoters, having three basic components: an enabling region, roughly 100 b.p. upstream the cap site; the stability profiles of enabling regions show some homology, but their important features appear complex; a GC-rich activation domain, 50-70 b.p. upstream the cap site; its anticipated role in enzyme activation is consistent with several lines of experimental data; althoughits high stability could provide firm support for attachement of some enzymatic element (a situation also encounterd in procaryotic promoters (53)), the precise activation mechanism remains obscure; an AT-rich trap domain (TATA box), centered 30 b.p. upstream the associated cap site; its low stability could ease melting-in of specific elements involved in the initiation of transcription. Activity feedback between these three domains is possible.

We observe that, in contrast to regulation by a fixed chemical structure, which allows a mere binary, all-or-none issue, regulation by a collective physical property like stability is much more flexible, as it allows to introduce many nuances between the two regulation extremes (full expression or full repression). Furthermore, since it is known that the stability profile of a given DNA segment depends on its environment, regulation by the modification of DNA stability induced by environment is envisageable (54).

The nature of the relevant physical properties involved in promoter definition and initiation of transcription can only be suspected at present, with some simple criteria (stability, GC or AT clusters), but may be much more subtle. A true understanding of the promoter action needs their full elucidation, both by in vivo studies of mutated target elements (reverse genetics) and their physical investigation by appropriate methods.

*To whom all correspondence should be addressed.

ACKNOWLEDGEMENTS.

We are specially indebted to Prof. P.Chambon for encouragement, advice and enlighting discussions. Expert assistance from the staff members of Paris Sud Informatique is acknowledged. We are kindly grateful to Fondation pour la Recherche Médicale for financial support. The model in Fig. 1 was skillfully built by Etienne and Christophe R..

REFERENCES.

1. Pribnow, D., (1975) J. Mol. Biol. 99, 419-443.
2. Goldberg, M.L., (1979) PhD. Thesis, Standford University. Palo Alto, California.
3. Shine, J., Dalgarno, L., (1974) Proc. Natl. Acad. Sci. USA 71, 1342.
4. Rosenberg, M., Court, D., (1979) Ann. Rev. Genet. 13, 319-353.
5. Scherer, G.F.E., Walkinshaw, M.D., Arnott,S., Morre, D.J., (1980) Nuc. Acid Res. 8, 3895.
6. Breathnach, R., Chambon, P., (1981) Ann. Rev. Biochem. 50, 349-383.
7. Ziff, E.R., Evans, M.R., (1978) Cell 15, 1463-1475.
8. Wells, R.D., Blakesly, R.W., Hardies, S.C., Horn, G.T., Larson, J.E., Selsing, E., Baird, J.F., Chan, H.W., Dodgson, J.B., Jensen, K.F., Nes, I.F., Wartell, R.M., (1977) CRC Critical Review of Biochemistry 4, 305-340.
9. Mardsen, M.P.F., Laemmli, U.K., (1979) Cell 17, 849-858.
10. Benham, C.J., (1981) J. Mol. Biol. 150, 43-68.
11. Gabarro-Arpa, J., Ehrlich, R., Rodier, F., Reiss, C., (1980) in DNA Recombination Interaction and Repair, Zadrazil, S., Sponar, J., eds., Pergamon Press New-York, 211-221.
12. Ehrlich, R., Marin, M., Gabarro-Arpa, J., Rodier, F., Schmitt, B., Reiss, C., (1981) C. R. Acad. Sci. Paris 292, 5-8.
13. Rodier, F., Gabarro-Arpa, J., Ehrlich, R., Reiss, C.,(1982) Nucl. Acids Res. 10, 391-402.
14. Benoist, C., Chambon, P., (1981) Nature 290, 304-310.
15. Moreau, P., Hen, R., Wasylyk, B., Evrett, R., Gaub, M. P., Chambon, P., (1981) Nucl. Acids Res. 9, 6047-6067.
16. McKnight, S. L., Kingsbury, R., (1982) Science 217, 316-324.
17. Falkow, S., Cowie, D. B., (1968) J. Bacteriol. 96, 777-795.
18. Yabuki, S., Wada, A., Uemura, K., (1969) J. Biochem. (Tokyo) 65, 443.
19. Vedenov, A.A., Dykhne, A.M., Frank-Kamenetskii, M.D., (1972) Sov. Phys. Usp. 14, 715.
20. Wada, A., Yabuki, S., Husimi, Y., (1980) Critic. Rev. Biochem. 9, 87-144.
21. Reiss, C., Gabarro-Arpa, J., (1977) in Prog. Molec. and Subcell. Biol., Hahn, F. E., ed., Spriger Ver. Berlin, Vol. 5, 1-31.
22. Studier, F. W., (1965) J. Mol. Biol. 41, 199-209.
23. Bloomfield, V. A., Crothers, D. M., Tinoco, I., (1974) Phys. Chem. of Nucl. Acids, Harper and Row, New York, Chap. 6.
24. Azbel, M. Ya., (1973) Phys. Rev. Letters, 31, 589-592.
25. Gabarro-Arpa, J., Michel, F., (1982) Biochimie, 64, 99-112.
26. Michel, F., Gabarro-Arpa, J., Dujon, B., (1982) Biochimie 64,113-126.
27. Gabarro-Arpa, J., Tougard, P., Reiss, C., (1979) Nature 280, 515-517.
28. Benight, A. S., Wartell, R. M., Howell, D. K., (1981) Nature 289, 203-205.

29. Levinson, B., Khoury, G., Van de Woude, G., Gruss, P., (1982) Nature 295, 568-572.
30. Vollenveider, H., Fiand, H., Szybalsky, W., (1979) Science 205, 508-510.
31. Grosschedel, R., Birnstiel, M. L., (1980) Proc. Natl. Acad. Sci. USA 77, 1432-1436.
32. Grosveld, G.C., de Boer, E., Shewmaker, C. K., Flavell, R. A., (1982) Nature 295, 120-125.
33. Grez, M., Land, H., Gieseke, K., Schutz, G., Jung, A., Sippel, A.E., (1981) Cell 25, 743.
34. Contreras, R., Fiers, W., (1981) Nucl. Acids Res. 9,215.
35. McKnight, S. L., Gavis, E. R., Kingsbury, R., (1981) Cell 25, 385-398.
36. Konkel, D. A., Maizel Jr., J. V., Leder, P., (1979) Cell 18, 865-873.
37. Nishioka, Y., Leder, P., (1979) Cell 18, 875-882.
38. Haynes, J. R., Rosteck Jr., P., Lingrel, J. B., (1980) Proc. Natl. Acad. Sci. USA 77, 7127-7131.
39. Gannon, F., O'Hare, K., Perrin, F., Le Pennec, J. P. , Benoist, C., Cochet, M., Breathnach, R., Royal, A., Garapin, A., Cami, B., Chambon, P., (1979) Nature 278, 429-434.
40. Van Ooyen, A., Van den Berg, J., Mantei, N., Weissmann, C., (1979) Science 206, 337-344.
41. Nagata, S., Mantei, N., Weissmann, C., (1980) Nature 287, 401-408.
42. Alestrom, P., Akusjarvi, G., Perricaudet, M., Mathews, M. B., Klessig, D. F., Pettersson, U., (1980) Cell 19, 671-681.
43. Akusjarvi, G., Pettersson, U., (1979) J. Mol. Biol. 134, 143-158.
44. Busslinger, M., et al., (1980) Nucl. Acids. Res. 8, 957.
45. Cochet, M. Gannon, F., Hen, R., Maroteaux, L., Perrin, F., Chambon, P., (1979) Nature 282, 567.
46. Torok, I., Karch, F., (1980) Nucl. Acids. Res. 8, 3105-3123.
47. Sures, I., Levy, S., Kedes, H. L.,(1980) Proc. Natl. Acad. Sci. USA 77, 1265-1269.
48. Dijkema, R., Dekker, B. M. M., Van Ormondt, H., (1980) Gene 9, 141-146.
49. Tyndall, C., La Mantia, G., Thackn, C. M., Favaloro, J., Kamen, R., (1981) Nucl. Acids. Res. 10, 6231.
50. Liebhaber, S. A., Goossens, M. J., Yuet Wai Kan, (1980) Proc. Natl. Acad. Sci. USA 77, 7054-7058.
51. Pelham, H. R. B., (1982) Cell 30, 517-528.
52. Grosveld, G. C., de Boer, E., Shewmaker, C. K., Flavell, R. A., (1982) Nature 295, 120-126.
53. Ehrlich, R., Marin, M., Gabarro-Arpa, J., Rodier, F., Schmitt, B., Reiss, C., (1981) C. R. Acad. Sci. Paris 292-III, 177-180.
54. Ehrlich, R., Rodier, F., Gabarro-Arpa, J., Marin, M., Schmitt, B., Reiss, C., (1981) C. R. Acad. Sci. Paris 293-III, 1-4.