Autoregulation of Expression of T4 Gene 32: a Quantitative Analysis

PETER H. VON HIPPEL,[1] STEPHEN C. KOWALCZYKOWSKI,[1]† NILS LONBERG,[1]‡ JOHN W. NEWPORT,[1]§ LELAND S. PAUL,[1] GARY D. STORMO,[2] AND LARRY GOLD[2]

Institute of Molecular Biology and Department of Chemistry, University of Oregon, Eugene, Oregon 97431,[1] and Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado 80309[2]

Gene expression, manifested as the orderly production of specific proteins of appropriate types and amounts in defined progressions, is regulated at virtually every step of mRNA synthesis and its translation into protein.

Exploration of the Jacob-Monod (1961) operon model, and its progressive modification and expansion as more complex patterns of control have been revealed experimentally, has demonstrated that differential regulation of the synthesis of particular mRNA molecules occurs at initiation, during elongation, and at termination of transcription. Control always involves some form of feedback, most simply through the metabolite-level-dependent binding of constitutively synthesized repressor or activator proteins to regulatory sites on the DNA (e.g., see Savageau, 1979). However, control may also be autogenous in nature in that such repressor or activator proteins may directly modulate the expression of their own structural genes, at either the transcriptional or the translational level (Goldberger, 1974; Savageau, 1979).

Steitz (1979) and Gold et al. (1981) have recently reviewed a large body of evidence that demonstrates that gene regulation at the initiation step of translation is quite general. In these systems the binding of control proteins or the presence of elements of mRNA secondary structure can modulate the access of ribosomes to ribosomal initiation sites and thus regulate the relative levels of expression of mRNA sequences coding for various proteins. In a particularly simple (in concept) version of such a system, the protein at issue binds specifically and reversibly to its own ribosomal initiation site ("translational operator") and thus controls its own synthesis. The autoregulation, by this mechanism, of the production of gene 32 protein in T4 phage infection represents the best-understood example of the operation of such a system and is described in this chapter.

In principle such control is particularly appropriate for the regulation of the free intracellular concentrations of proteins required in considerable quantity as structural elements of multiprotein "organelles" such as DNA replication complexes, ribosomes, etc. The structural protein is produced as needed and is incorporated into the organelle, typically by a self-assembly process based on coupled equilibria. When the organelle has been saturated with the protein at issue, further synthesis leads to an increase in the free intracellular concentration of this species. Ultimately a critical concentration is passed, and the protein binds to a regulatory site on its own mRNA, leading to reversible shutoff of synthesis. Economy of protein design suggests that this latter binding should involve generally the same interactions, and the same binding site(s), as are used in the functional binding of the protein as a structural component.

Clearly, while simple in concept, this scenario can lead to difficulties if the binding of the protein is relatively nonspecific, since (for example) this can result in uncontrolled binding to, and perhaps premature shutoff of, initiation sites for unrelated proteins as well. The necessary discrimination involves finely tuned systems of coupled binding equilibria. The gene 32 protein system is particularly well suited for demonstrating the possibilities and problems inherent in such control mechanisms. A similar treatment will be presented elsewhere, in which the principles of the gene 32 protein autoregulation system are also extended to show how these same approaches might be used to explain the autoregulation of synthesis of coordinately regulated *systems* of proteins, such as those involved in the structure and assembly of the ribosome (Fairfield and von Hippel, manuscript in preparation).

AUTOREGULATION OF T4-CODED GENE 32 PROTEIN SYNTHESIS

Gene 32 protein is an essential component of the T4 DNA replication, recombination, and repair systems (for recent reviews, see Doherty et al., *in* J. F. Kane, ed., *Multifunctional Proteins*, in press; Williams and Konigsberg, this volume). It plays a "structural" (as opposed to a catalytic) role, binding in saturating amounts to the single-stranded DNA (ssDNA) that is transiently produced in the essential intermediate stages of these processes. Genetic and biochemical studies have shown that the total amount of gene 32 protein produced in a phage infection depends directly on the amount of intracellular ssDNA present (Gold et al., 1976; Krisch et al., 1974). It has also been shown that the synthesis of gene 32 protein is regulated at the translational level (Lemaire et al., 1978; Russel et al., 1976).

In effect, intracellular control of the free concentration of gene 32 protein involves an orderly progression of binding events. All ssDNA sequences are saturated as the level of free protein increases initially. Only after this process is complete does the free intracellular protein concentration rise to a threshold level high enough to permit binding to the gene 32 mRNA "translational operator" site (Russel et al., 1976), resulting in the specific cessation ("repression") of gene 32 protein synthesis. In vitro experiments have shown that this level of free protein concentration is

† Present address: Department of Molecular Biology, Northwestern School of Medicine, Chicago, IL 60611.

‡ Present address: Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, MA 02138.

§ Present address: Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94143.

not sufficient to permit binding to translational initiation sites of other T4 mRNAs (Lemaire et al., 1978), to permit binding to the very large reservoir of double-stranded DNA present in the cell (Jensen et al., 1976; Newport et al., 1981), or to prevent the reannealing of double-stranded DNA after the replication process is complete (von Hippel et al., 1982).

A combination of biochemical (Lemaire et al., 1978) and physical chemical (Jensen et al., 1976; Kelly et al., 1976; Kowalczykowski, Lonberg, Newport, and von Hippel, 1981; Newport et al., 1981) experiments has provided the necessary data for a quantitative molecular description of the autoregulatory cycle responsible for the establishment and maintenance of physiological levels of gene 32 protein in T4 infection of *Escherichia coli*. These studies are summarized briefly below and are described in full detail by von Hippel et al. (1982).

Binding Parameters for Gene 32 Protein

The binding of a protein to a nucleic acid lattice can be described by three thermodynamic constants (McGhee and von Hippel, 1974): the binding-site size (n; in units of nucleotide residues covered per protein monomer bound), the intrinsic association constant (K; in units of M^{-1}), and the cooperativity parameter (ω; unitless). These parameters have been measured for the binding of gene 32 protein to a variety of single-stranded deoxyribose- and ribose-containing homo- and heteropolynucleotides as a function of salt concentration and temperature (Kowalczykowski et al., 1981; Newport et al., 1981). The results show that n is constant at 7 (±1) nucleotide residues, that ω is constant at $\sim 2 \times 10^3$, and that K varies with nucleotide composition of the lattice, salt concentration, and temperature. These measurements have permitted us to calculate values of the effective affinity constant of gene 32 protein binding in the cooperative polynucleotide binding mode ($K\omega$) to ssDNA and RNA sequences either of known sequence or of average T4 DNA composition, under physiological conditions. We define these conditions as a temperature of 37°C and a salt concentration of 0.23 M NaCl; this salt concentration has been shown to be approximately equivalent, in terms of the strength of protein-nucleic acid binding interactions, to the actual intracellular ionic environment (Kao-Huang et al., 1977).

In Vitro Repression Experiments

Lemaire et al. (1978) have conducted experiments that demonstrate the translational repression of gene 32 protein in vitro, using a cell-free translation system containing a crude RNA preparation from T4-infected *E. coli* cells, and ribosomes, tRNA, and supernatant proteins derived from uninfected *E. coli*. The results of these experiments may be summarized as follows. (i) Gene 32 protein binds preferentially to a specific component of the RNA derived from T4-infected cells. Since shutoff is specific for the synthesis of gene 32 protein, this component must be a portion of the gene 32 mRNA. (ii) The abruptness with which shutoff occurs as a function of added gene 32 protein suggests that this repression (and the binding of the protein to the gene 32 mRNA that is assumed to be responsible for it) is cooperative in gene 32 protein concentration. (iii) ssDNA effectively binds gene 32 protein more tightly than does either ssRNA in general or the gene 32 mRNA translational operator site. (iv) The binding affinity of gene 32 protein for the gene 32 mRNA operator is larger than that for most other RNA constituents in the system and is comparable to that of (unstructured) poly(rU). (v) Double-stranded DNA, and also the other components of the cell-free translation system, bind gene 32 protein less strongly than does the gene 32 mRNA operator. (vi) The addition of gene 32 protein to levels that are three- to fourfold greater than required to halt gene 32 protein synthesis does shut off the synthesis of other T4 proteins in the cell-free translation system, suggesting that the gene 32 mRNA operator site differs only quantitatively (in terms of gene 32 protein binding) from translational control sites on other T4 mRNAs. These and other data can also be used to estimate that the free intracellular gene 32 protein concentration maintained in vivo (during T4 infection) is ~3 μM (von Hippel et al., 1982).

Calculation of In Vivo Gene 32 Protein Binding (Titration) Curves for Various Structured and Unstructured Nucleic Acid Targets

By using the known binding parameters for gene 32 protein to various nucleic acid sequences, titration curves for the binding of gene 32 protein to various potential nucleic acid targets under physiological conditions have been calculated (von Hippel et al., 1982). The results are fully and quantitatively compatible with the experimental facts outlined above and, together with the sequencing data of Krisch and coworkers (Krisch and Allet, 1982; Krisch et al., 1980), have permitted the definition of the gene 32 mRNA translational operator site.

A two-state calculation was used initially to determine the expected levels of binding of gene 32 protein to unstructured ssDNA and RNA lattices. The results showed that long ssDNA lattices of average T4 composition, unencumbered by secondary structure, would be expected to saturate at ~0.01 μM free gene 32 protein, whereas comparable RNA lattices would saturate at ~0.3 μM protein. Both types of lattice should thus be fully saturated at physiological gene 32 protein concentrations.

However, most nucleic acid sequences in the cell are partially or completely involved in secondary structure. As a consequence, the favorable (to binding) free-energy change (ΔG_{bind}) involved in the interaction of gene 32 protein with single-stranded lattices will be opposed by the conformational free energy ($\Delta G^\circ_{\text{conf}}$) favoring the maintenance of partially double-stranded structures. This conformational free energy can be estimated by using the approach and parameters developed by Tinoco et al. (1973). As a consequence, higher free gene 32 protein concentrations are needed to saturate such initially structured nucleic acid lattices. Such calculations reveal (data not shown) that, because of its tighter binding to ssDNA lattices, physiological concentrations of gene 32 protein will saturate DNA lattices containing stem-loop structures with as much as 70% of the sequences involved in base pairing. Thus, virtually all secondary structure that might be expected to develop adventitiously in single-stranded regions during DNA replication should be removable by gene 32 protein at the controlled free in vivo concentration.

The situation for mRNA should be quite different. A variety of lines of evidence (see Gold et al. [1981] for a summary) suggests that mRNA secondary structure is crucial for biological activity and thus should not be "melted" by gene 32 protein. The calculated results (Fig. 1) are fully compatible with this expectation, showing that because of the lesser (relative to ssDNA) affinity of gene 32 protein for ssRNA, only very weak elements of mRNA secondary structure should be melted at physiological gene 32 protein concentrations.

Finite Nucleic Acid Lattice Effects

To this point, the calculations described above were carried out by using a two-state "infinite lattice" model. In this model it is assumed that the stem-loop regions (for example) for which binding curves are being calculated are already flanked by gene 32 protein-complexed sites. This means that every protein monomer bound will contribute a full "unit" of both intrinsic binding affinity (K) and binding cooperativity (ω) to the interaction. Thus

$$\theta = \frac{(K_{\text{conf}})(K_{\text{bind}})[\text{P}]^m}{1 + (K_{\text{conf}})(K_{\text{bind}})[\text{P}]^m} \qquad (1)$$

where θ = the fraction of the lattice sites under consideration that have been saturated at free protein concentration [P], m = the length of the lattice sequence under consideration in protein monomer units ($m = N/n$; where n = the protein site size and N = the lattice segment length in nucleotide residues), K_{conf} = $[\text{NA}_{\text{ss}}]/[\text{NA}_{\text{ds}}]$ ($[\text{NA}_{\text{ss}}]$ and $[\text{NA}_{\text{ds}}]$ represent, respectively, the molar concentrations of open [single-stranded] and duplex [base-paired] nucleic acid lattice, in units of nucleotide residues), and

$$K_{\text{bind}} = (K\omega)_1(K\omega)_2 \ldots (K\omega)_m = \prod_{i=1}^{i=m} (K\omega)_i \qquad (2)$$

We note that $K_{\text{bind}} = (K\omega)^m$ for infinite lattices of constant composition.

This model is quite appropriate for considering the titration by gene 32 protein of an mRNA segment containing a weak stem-loop structure (hairpin) or for "filling in" a single-stranded lattice segment comprising the transient "single-stranded window" in a moving DNA replication fork, but it is less valid for estimating the degree of saturation (within an mRNA molecule) of single-stranded regions that are flanked by elements of secondary structure too stable to be melted at the physiological gene 32 protein concentration. For such regions a finite lattice calculation needs to be made, where

$$K_{\text{bind}} = K_1(K\omega)_2(K\omega)_3 \ldots (K\omega)_m = K_1 \prod_{i=2}^{i=m} (K\omega)_i \qquad (3)$$

We note that the finite lattice binding definition of K_{bind} (equation 3) differs from that for infinite lattice binding (equation 2) only by the loss of one "unit of ω", but for short sequences this loss can make an enormous difference in the resulting titration curve (Fig. 2). Figure 2 thus shows that, due to this finite-lattice effect, even totally unstructured mRNA sequences (of average T4 composition) do not bind gene

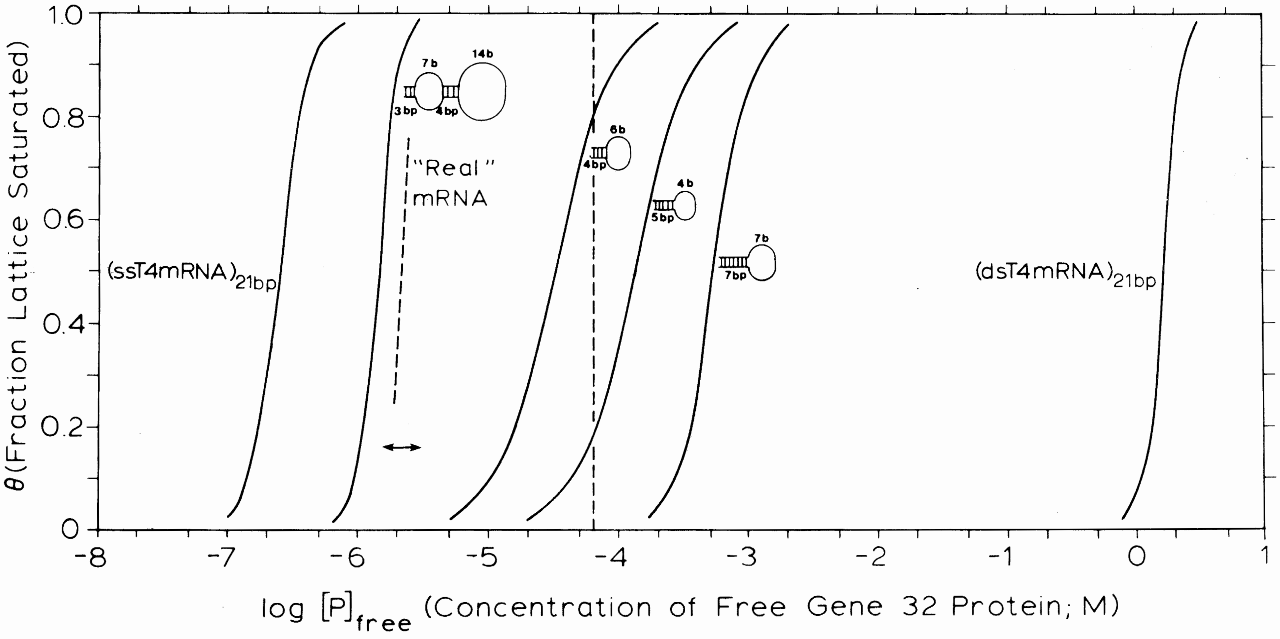


FIG. 1.–Binding curves for the melting and complexation by gene 32 protein of various hypothetical initially looped and bulged T4 mRNA structures, plotted as a function of free gene 32 protein. The titration curves correspond, respectively, to the indicated stem-loop (and/or bulge) structures. The sloped dashed line labeled "real mRNA" is the approximate binding isotherm for the gene 32 mRNA control site, as estimated from the experiments of Lemaire et al. (1978). (Figure from von Hippel et al., 1982.)

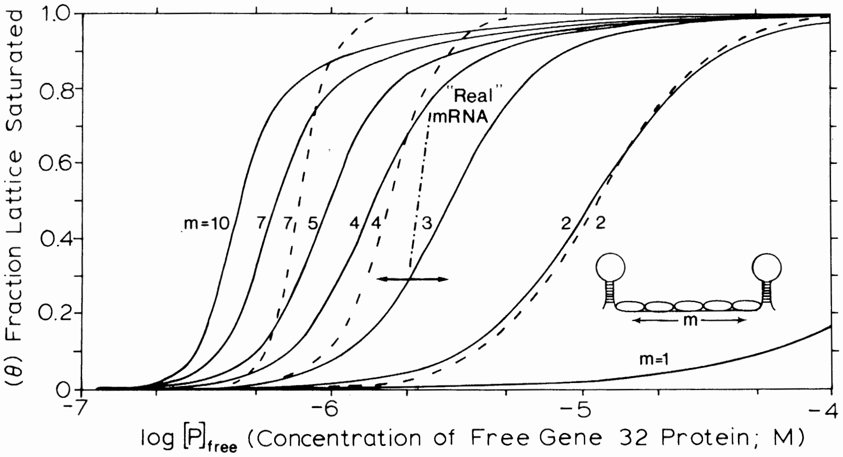


FIG. 2. Binding curves for the finite mRNA lattices of varying length. The dashed curves represent the two-state approximation, calculated as outlined in the text. The solid curves were calculated by the "exact" method of Epstein (1978); for further details see Newport et al. (1981). The lengths of the lattices are defined in units (m) of protein monomer binding sites. The site size of gene 32 protein binding cooperatively in the polynucleotide binding mode is seven nucleotide residues. Thus, the lengths of the respective finite lattices, in units of nucleotide residues, are $7m$. (Figure from von Hippel et al., 1982.)

32 protein under physiological conditions and protein concentrations if they have a lattice length (m) of less than 4 (~28 nucleotide residues). Furthermore, also due to this effect, even longer regions containing elements of weak secondary structure will remain uncomplexed. We expect that under physiological conditions the average mRNA molecule will be highly structured; thus, sequences that are sufficiently unstructured to bind gene 32 protein under intracellular conditions may be relatively rare.

The Gene 32 mRNA Translational Operator Site

The calculations above suggest that, in principle, the simplest way to define the gene 32 protein translational operator site, and to ensure that it saturates at lower free gene 32 protein concentrations than do control sequences on other T4 mRNAs, is to have the gene 32 mRNA operator consist of a uniquely unstructured segment, as originally proposed by Russel et al. (1976). The combination of the availability of $K\omega$ values for all of the relevant nucleic acid lattices, the recent sequencing of gene 32 mRNA by Krisch et al. (1980) and Krisch and Allet (1982), and the availability of a large T4 DNA sequence library (Gold et al., 1981; Schneider et al., 1982; Stormo and Schneider, unpublished data) now makes it possible to test this suggestion quantitatively.

The sequence surrounding the initiation codon of the gene 32 message is shown in Fig. 3. In mRNA sequences this region contains the information for translational initiation, and thus this general sequence clearly is the most logical candidate for the gene 32 mRNA translational operator site. This view is based on the simplest translational repression model, in which gene 32 protein (as repressor) competes with the ribosome for this operator-initiator site.

The sequence of gene 32 mRNA in the vicinity of the initiation codon is remarkable, even for a phage containing 66% adenine-plus-thymine residues. As Fig. 3 shows, the ribosome-binding site region contains a stretch of 40 nucleotides (residues 33 to 72, inclusive) in which the only bases other than A or U are the three nearly essential G residues of the Shine-Dalgarno sequence and the initiation codon (Gold et al., 1981). Values of $\Delta G^{\circ}_{\text{conf}}$ have been computed for a variety of arbitrary segments within the gene 32 initiation sequence to determine whether an unstructured domain of sufficient length to serve as an operator site could exist in this region within the quantitative constraints outlined above. Some of the results are shown in Fig. 3. In essence, it was found that the longest (unstructured and partially structured) potential operator sequence that can be saturated under intracellular conditions and at the regulated gene 32 protein concentration is represented by line D in Fig. 3. This sequence is shown in the bound conformation (complete with stable flanking hairpins) at the bottom of the figure; it binds nine gene 32 protein monomers!

Is the Gene 32 mRNA Operator Sequence Unique?

It was also, of course, necessary to determine whether the proposed gene 32 mRNA operator sequence defined in Fig. 3 is unique. To this end calculations were carried out using the entire catalog of T4 nucleic acid sequences. The results showed that the proposed gene 32 mRNA operator has much less secondary structure than virtually any other sequences within the T4 sequence catalog (~5% of the total T4 genome). A comparison with more than 10 other T4 ribosome-binding sites showed none to be as unstructured as the proposed gene 32 mRNA operator (von Hippel et al., 1982).

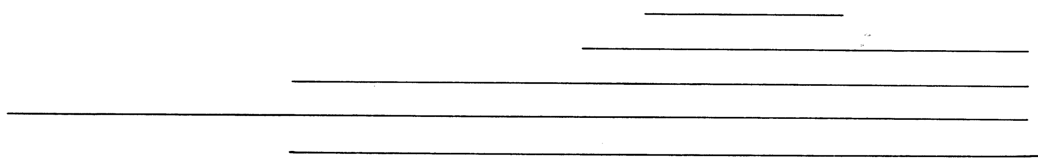
The T4 Gene 32 mRNA Autogenous Regulatory System

The conclusions outlined above are summarized in Fig. 4 for the actual T4 system. Figure 4 shows, as required, that the actual ssDNA sequences of the T4 DNA replication complex (and presumably also those for T4 DNA recombination and repair systems) are

```
                                                        met   lys   lys   thr   glu   ala   gln   ala   leu   gly (etc.)
                                                           phe   arg   ser   ala   leu   ala   met   lys   asn

          1         2         3         4         5         6         7         8         9         0         1         2         3         4         5
GCTCATGAGGTAAAGTGTCATAGCACCAACTGTTAATTAAATTAAATTAAAAAGGAAATAAAAATGTTTAAACGTAAATCTACTGCTGAACTCGCTGCACAAATGGCTAAACTGAATGGCAATAAAGGTTTTTCTTCTGAAGATAAAGGCGAGT
aaa bbbb          bbbbaaaccd  ee ffffff       ffffff   eed                       cc               ggggggghhh ii                      iihhh   jjj        jjj                  gggggg

   ΔG°conf(a-b) = -5.2                     ΔG°conf(c-f) = -3.6                                        ΔG°conf(g-j) = -14.6
```



line	nucleotide residues(N)	protein monomers(m)	ΔG°_{conf}
B	18	2	0
C	39	5	-2.4
D	65	9	-3.6
E	89	12	-8.8
F	130	18	-18.2

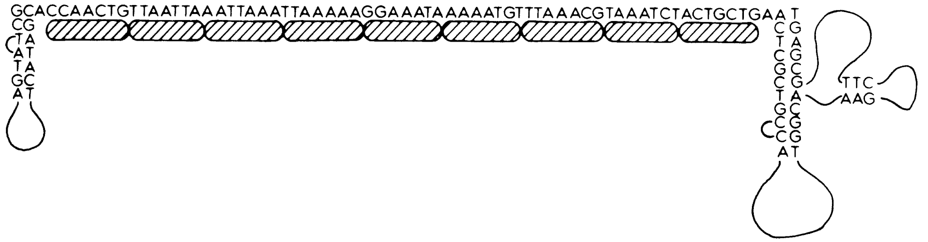

GCACCAACTGTTAATTAAATTAAATTAAAAAGGAAATAAAAATGTTTAAACGTAAATCTACTGCTGAA
CG
TA
AT
TA
GC
AT
T
CG
TA
CG
GC
CG
TA
GC
CG
CG
AT
TTC
AAG

FIG. 3. Sequence and conformational stability of the putative gene 32 mRNA operator site and vicinity. At the top is the DNA sequence (noncoding strand only; the sequence as written corresponds to mRNA when T is replaced by U), with the beginning of the gene 32 protein sequence written above. The lower-case letters below the DNA sequence correspond to possible base-pairing interactions; i.e., the bases marked *aaa* can pair with the subsequent *aaa* sequence to form the stem of a hairpin structure, etc. The lines (labeled B through F) correspond to the segments tested as potential operator sites (see text). The structure at the bottom is the preferred operator sequence, drawn in a gene 32 protein-saturated conformation showing the proposed flanking hairpin termini. (Figure from von Hippel et al., 1982.)

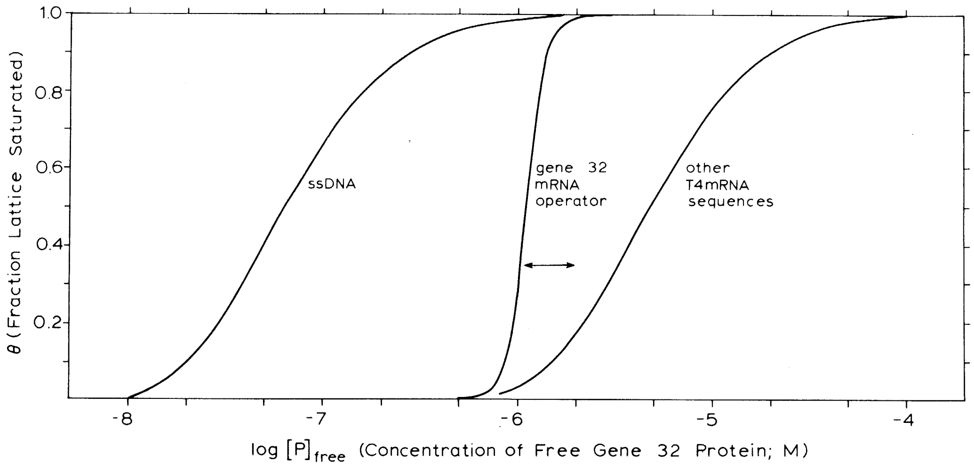


FIG. 4. Binding curves summarizing the gene 32 protein autoregulatory system. The ssDNA curve is calculated by using the real T4 DNA sequences with 50-residue lattice length (*N*) replication window and the infinite-lattice calculation mode. The gene 32 mRNA operator curve is calculated for the putative operator structure (Fig. 3, line D) shown at the bottom of Fig. 3. The "other mRNA" curve is calculated by using real T4 sequences with $N = 50$ and the finite-lattice approach. (Figure from von Hippel et al., 1982.)

saturated with gene 32 protein at concentrations well below the autoregulated value. The proposed gene 32 mRNA translational operator site then saturates quite sharply (cooperatively) at free protein concentrations just below the autoregulated level. As also required, other T4 mRNA initiation (ribosomal binding) sequences are not appreciably complexed at the maintained intracellular free gene 32 protein concentration.

SUMMARY AND OUTLOOK

The results presented here show that the regulation of expression of gene 32 of phage T4 can be modeled, using physical chemical binding data, to provide a quantitative and functionally economical picture of this system that is fully consistent with available biochemical and genetic information. The same approach, suitably modified to base control on "heteroprotein" (rather than on "homoprotein") cooperativity, appears to provide a useful way of thinking about the assembly and the autoregulation of synthesis of the components of the *E. coli* ribosome (Fairfield and von Hippel, manuscript in preparation; von Hippel and Fairfield, 1982) and, perhaps, the T4 replisome (Campbell and Gold, 1982). As further quantitative information is obtained about other T4 proteins, it may turn out that related approaches will help to explain their regulation as well (see, for example, the *regA* system [Karam et al., 1981; Karam and Wiberg, this volume]). However, it is already clear that the gene 32 system, per se, represents the simplest possible prototype; the control of expression of most of the other T4 genes will probably be much more interrelated and thus much more complex in quantitative detail.

This work was supported in part by Public Health Service research grants GM-15792 and GM-29158 (to P.H.v.H.) and GM-19963 (to L.G.) from the National Institutes of Health. J.W.N. and L.S.P. were predoctoral trainees on Public Health Service training grants GM-07759 and GM-70015.

We are very grateful to Henry Krisch and his colleagues for making the DNA sequence of the gene 32 region of phage T4 available to us before publication.