*J. Mol. Biol.* (1982) **162**, 795–818

# Autoregulation of Gene Expression

## Quantitative Evaluation of the Expression and Function of the Bacteriophage T4 Gene *32* (Single-stranded DNA Binding) Protein System

PETER H. VON HIPPEL, STEPHEN C. KOWALCZYKOWSKI†, NILS LONBERG‡
JOHN W. NEWPORT§, LELAND S. PAUL

*Institute of Molecular Biology and Department of Chemistry*
*University of Oregon*
*Eugene, Ore. 97431, U.S.A.*

GARY D. STORMO AND LARRY GOLD

*Department of Molecular, Cellular and Developmental Biology*
*University of Colorado*
*Boulder, Col. 80309, U.S.A.*

*(Received 3 May 1982)*

The free concentration of bacteriophage T4-coded gene *32* (single-stranded DNA binding) protein in the cell is autoregulated at the translational level during T4 infection of *Escherichia coli*. The control of the synthesis of this protein reflects the following progression of net (co-operative) binding affinities for the various potential nucleic acid binding targets present: single-stranded DNA > gene *32* mRNA > other T4 mRNAs ≥ double-stranded DNA. In this paper we show that the free concentration of gene *32* protein is maintained at 2 to 3 $\mu$M, and use the measured binding parameters for gene *32* protein, extrapolated to intracellular conditions, to provide a quantitative molecular interpretation of this system of control of gene expression. These results are then further utilized to define the specific autoregulatory binding sequence (translational operator site) on the gene *32* mRNA as a uniquely unstructured finite binding lattice terminated by elements of secondary structure not subject to melting by gene *32* protein at the autoregulated concentration, and to predict how this site must differ from those found on other T4 messenger RNAs. It is shown that these predictions are fully consistent with available T4 DNA sequence data. The control of free protein concentration as a method of genome regulation is discussed in terms of other systems to which these approaches may apply.

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

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

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



0022-2836/82/360795-24 $03.00/0 

## 1. Introduction

The elucidation of molecular mechanisms of the regulation of gene expression continues to be one of the central preoccupations of molecular biologists. The tight binding of genome regulatory proteins to unique chromosomal target sequences (e.g. repressors to operator sites) comprises the central element of one important type of control system. Such binding derives its specificity from the interaction of complementary matrices of hydrogen bond donors and acceptors located, respectively, in the binding site of the protein and in the grooves of the specific double-stranded DNA base-pair sequence (e.g. see von Hippel, 1979).

Recently, however, mechanisms based on completely different physical chemical principles have begun to emerge as well. For example, regulatory systems of striking specificity can be developed by utilizing protein binding co-operativity to amplify rather modest differences in intrinsic binding affinities. When combined with a definitive feedback mechanism for holding free protein concentration at a fixed concentration, such systems can fully saturate certain specific nucleic acid binding targets while leaving others totally uncomplexed. The autogenous regulation of synthesis of the bacteriophage T4 gene *32* protein represents the simplest system of this sort, which has now been worked out in quantitative detail. We present here the approach and the important results for this case, and indicate how the same notions might be applied to the interpretation of other genome-regulatory mechanisms of greater complexity.

## 2. Materials and Methods

### (a) *Measurements of thermodynamic parameters*

Values of $n$, $K$ and $\omega$ (and $K\omega$) were determined (or calculated) for the binding of gene *32* protein to various natural and synthetic nucleic acid lattices as described elsewhere (Kowalczykowski *et al.*, 1981; Newport *et al.*, 1981*b*).

### (b) *Computer computations*

The program used to predict nucleic acid secondary structure (contributed by Dr Eugene Myers, University of Colorado) is similar to those reported by others (Nussinov & Jacobson, 1980; Zuker & Stiegler, 1981). It uses a dynamic programming algorithm to compute (in $N^2$ space and $N^3$ time, where $N$ is the length of the lattice in nucleotide residues) the thermodynamically most favorable structure for a given single-stranded sequence, using the usual folding rules involving base stacking, loop destabilization free energies, etc. (Tinoco *et al.*, 1973). We have not attempted to alter these rules to account for possible temperature or salt concentration effects on nucleic acid stability, nor have we considered possible DNA *versus* RNA stability differences. The bacteriophage T4 sequences examined come from a library (Schneider *et al.*, 1982) that now contains sequences totalling 6797 nucleotide residues of T4 in 5 fragments.

The stabilities of "local secondary structures" that may form in short stretches of DNA or RNA were calculated as follows. For each lattice we first calculated the most stable "folded" structure for the first $N$ residues ($N=30$, 40 or 50 residues and comprises the moving "window"). The window was then shifted $M$ residues in the 3′ direction ($M=10$ for $N=40$; and $M=20$ for $N=30$ and 50 residues), and the calculation was repeated. This process was continued until the last residue of the entire sequence appeared in the lattice window. This procedure was used to generate $\Delta G^0_{\text{conf}}$ values for 340, 687 and 335 (for $N=30$, 40 and 50 residue lattices, respectively) different, but overlapping, short sequences. The results are plotted in Fig. 6.

## 3. Background and Parameters

### (a) *Autoregulation of gene* 32 *protein synthesis*

The gene *32* protein of T4 plays an essential role in the life-cycle of this phage, participating in T4 DNA replication, recombination and repair (Doherty *et al.*, 1982). Genetic and biochemical studies (Krisch *et al.*, 1974; Gold *et al.*, 1976) have demonstrated that the total amount of gene *32* protein produced in a phage infection depends on the amount of intracellular single-stranded DNA present. Furthermore, in a series of studies *in vivo* and *in vitro* (Russel *et al.*, 1976; Lemaire *et al.*, 1978), it has been shown also that the synthesis of gene *32* protein is regulated at the translational level. Thus, after saturation of available single-stranded DNA lattices, the intracellular "pool" of free protein rises to a critical concentration and the synthesis is (reversibly) shut-off. Repression appears to be a consequence of the specific binding of the protein itself to a control region of the gene *32* messenger RNA; this control region has been called a "translational operator" (see Russel *et al.*, 1976; Karam *et al.*, 1981). Considerably higher concentrations of free protein are required to shut-off synthesis of other T4-coded proteins (Lemaire *et al.*, 1978), and to bind to the great excess of double-stranded DNA present in the cell (Jensen *et al.*, 1976; Newport *et al.*, 1981*b*).

In effect, intracellular control of the free concentration of gene *32* protein involves an orderly progression of binding events (Russel *et al.*, 1976; Lemaire *et al.*, 1978). As the concentration of free protein increases, all transiently present single-stranded DNA sequences are saturated first. 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 operator site, resulting in the specific cessation of the synthesis of this protein. Thus levels of free protein concentration sufficient to permit binding to translational initiation sites of other T4 mRNAs (and thus to inhibit the translation of other T4 gene products), or to bind to the very large reservoir of double-stranded DNA present, are not achieved under regulated conditions.

### (b) *Binding of gene* 32 *protein to various nucleic acid lattices*

The binding of a protein to a nucleic acid lattice is described by three thermodynamic parameters: the binding site size ($n$; in units of nucleotide residues per protein monomer); the intrinsic binding constant ($K$; in units of $\text{M}^{-1}$), and the co-operativity parameter ($\omega$; unitless). (See McGhee & von Hippel (1974) for further discussion of the definitions and measurement of these constants.) These parameters have been measured for the co-operative binding of gene *32* protein to various nucleic acid lattices. The results are (mostly) described in detail elsewhere (Jensen *et al.*, 1976; Kelly *et al.*, 1976; Kowalczykowski *et al.*, 1981; Newport *et al.*, 1981*b*; Lonberg *et al.*, 1981); a few additional measurements are reported here as well (see Table 1).

The site size $n$ for this binding is constant at 7 ($\pm 1$) nucleotide residues; the co-operativity parameter $\omega$ is also constant at $\sim 2 \times 10^3$ for gene *32* protein binding to various polynucleotides over a range of salt concentrations. The intrinsic binding constant $K$ of the protein to the lattice varies with nucleotide composition (sugar and base type), temperature and salt concentration.

Experiments with many polynucleotide lattices have shown that the standard binding free energy of a gene *32* protein monomer to any particular natural DNA or RNA lattice can be calculated as the compositionally weighted average of the binding free energies of the protein for the appropriate (deoxyribo- or ribo-) homopolynucleotides (Newport *et al.*, 1981*a*,*b*). Thus:

$$\Delta G^0_{\text{bind}} = \sum_i f_i (\Delta G^0_{\text{bind}})_i, \tag{1}$$

where $f_i$ is the fraction of the total nucleotide content of the sequence represented by nucleotide residue $i$, and $(\Delta G^0_{\text{bind}})_i$ is the standard free energy change for the binding of the protein to a homopolynucleotide lattice of type $i$. In terms of $K\omega$:

$$(K\omega)_{\text{bind}} = \prod_i (K\omega)_i^{f_i}, \tag{2}$$

where $(K\omega)_{\text{bind}}$ is the observed net co-operative binding affinity for the particular DNA or RNA lattice, and $(K\omega)_i$ are the equivalent parameters for the component homopolynucleotide lattices under the same conditions†.

### (c) *Binding affinities under physiological conditions*

For the purposes of this paper, we must establish a set of $K\omega$ values that apply to the binding of gene *32* protein to various DNA and RNA sequences in the infected *Escherichia coli* cell under physiological conditions. To this end, we must first define "physiological" temperature and salt concentration, since the co-operative binding of gene *32* protein to single-stranded nucleic acids is somewhat temperature dependent (Kowalczykowski *et al.*, 1981) and highly dependent on salt concentration and salt type (Kowalczykowski *et al.*, 1981; Newport *et al.*, 1981*a*,*b*).

For physiological temperature we use 37°C, both because most laboratory infections of *E. coli* by T4 are conducted at this temperature and because Lemaire *et al.* (1978) carried out their *in vitro* experiments on the translational repression of gene *32* protein synthesis at 37°C. We use 0·23 M-NaCl as equivalent, in terms of the strength of protein–nucleic acid binding interactions, to the intracellular salt concentration. This value was determined using an *E. coli* minicell mutant to measure the ratio of *lac* repressor free in the cell to that bound non-specifically to the chromosomal DNA (Kao-Huang *et al.*, 1977; D. W. Noble & P. H. von Hippel, unpublished results). This established an *in vivo* binding constant ($K_{\text{RD}}$) for *lac* repressor to non-operator DNA and, since the salt concentration dependence of the binding of *lac* repressor to non-specific DNA is known *in vitro* (deHaseth *et al.*, 1977; Revzin & von Hippel, 1977), a salt concentration equivalent to that of the effective intracellular ionic environment could be determined.

Values of $K\omega$ for gene *32* protein binding to various polynucleotide lattices at 37°C in the presence of 0·23 M-NaCl have been calculated and are summarized in Table 1. Most of our actual measurements of $K\omega$ were conducted at 20 to 25°C (Kowalczykowski *et al.*, 1981; Newport *et al.*, 1981*b*). The enthalpy change for this binding has been measured with polyriboethenoadenylic acid (poly(r$\varepsilon$A)); an average $\Delta H^0_{\text{bind}} \simeq -22\ (\pm 2)$ kcal/mol was established at several salt concentrations.

† Further details of the conditions under which eqns (1) and (2) were derived are given by Newport *et al.* (1981*b*). We note that eqn (1) in that publication was printed incorrectly; it should read as eqn (2) here.

This corresponds to an approximately fourfold decrease in $K\omega$ in going from 25 to 37°C. Direct measurements of $K\omega$ for the binding of gene *32* protein to T4 DNA and poly(rU) at 25 and 37°C roughly confirm this value, and support our earlier suggestion (Kowalczykowski *et al.*, 1981) that $\Delta H^0_{\text{bind}}$ measured with poly(rεA) can be generally applied to the binding of this protein to other polynucleotide lattices. Thus, except for those measured directly in this study, the $K\omega$ values in Table 1 have been corrected to 37°C, as indicated above, and extrapolated to 0·23 M-NaCl as described elsewhere (see Fig. 4 and Table 2 of Newport *et al.*, 1981*b*).

Values of $K\omega$ for single-stranded T4 DNA and T4 mRNA (assumed to have the same average base composition as T4 DNA) have been calculated from the homopolynucleotide binding data using equation (2). The validity of equation (2) for making such calculations has been confirmed (Newport *et al.*, 1981*b*) by showing that the same values of $K\omega$ were obtained for single-stranded bacteriophage $\phi$X174 DNA by direct measurement and by calculation using equation (2) and the average base composition of the DNA. In Table 1 we show that such agreement is also obtained between measured and calculated values of $K\omega$ for single-stranded (denatured) T4 DNA. In addition to further validating equation (2) as a means of calculating $K\omega$, this last result also reaffirms that the expected binding affinity of gene *32* protein for single-stranded T4 DNA is not significantly altered by the substitution of glucosylated hydroxymethylcytosine for cytosine residues in this DNA.

### (d) *Repression experiments* in vitro

Lemaire *et al.* (1978) have conducted experiments on the translational repression of gene *32* protein synthesis *in vitro*. Using a crude RNA preparation from T4-infected *E. coli* cells, together with a cell-free translation system consisting of

TABLE 1

*Values of* K$\omega$ *for the co-operative binding of gene* 32 *protein to single-stranded polynucleotides at 37°C and 0·23* M-*NaCl*

| Polynucleotide | $K\omega$ ($M^{-1}$)[a] |
|---|---|
| Poly(rC)[b] | $3\times10^4$ |
| Poly(rU)[b, c] | $4\times10^6$ |
| Poly(rA)[b] | $3\times10^6$ |
| Poly(rG) (est.)[b, d] | ($\sim10^8$) |
| T4 mRNA (average base composition)[e] | $\sim4\times10^6$ (calc.) |
| Poly(dC)[b] | $2\times10^8$ |
| Poly(dU)[b, f] | $5\times10^7$ |
| Poly(dA)[b] | $2\times10^7$ |
| Poly(dG) (est.)[b, d] | ($\sim10^9$) |
| T4 DNA (average base composition)[e] | $\sim10^8$ (calc.) |
| T4 DNA[c] | $8\times10^7$ |

[a] $K\omega$ is calculated per gene *32* protein monomer (binding co-operatively with a site size of 7).
[b] Extrapolated from the data of Newport *et al.* (1981*b*) as described in the text.
[c] Extrapolated from measurements made in this study.
[d] Corresponds to $K\omega$ values for isolated rG or dG residues in a natural RNA or DNA.
[e] 34% G·C.
[f] Corresponds to $K\omega$ values for isolated dT residues in a natural DNA sequence.

ribosomes, tRNA and supernatant proteins derived from uninfected *E. coli*, they have shown that the rate of synthesis of an amber (inactive) fragment of gene *32* protein is independent of the amount of active gene *32* protein added to the system, up to a total concentration of $\sim 3$ $\mu$M. The length of the protein concentration-independent "plateau" region in a plot of rate of fragment synthesized *versus* amount of total gene *32* protein added can be extended by the addition of single-stranded DNA.

With further increases in total gene *32* protein added (beyond $\sim 3$ $\mu$M-protein in the absence of added ssDNA†), the rate of synthesis of the gene *32* protein amber fragment decreases abruptly (and reversibly), falling to less than 10% of the plateau level at $\sim 4$ $\mu$M-protein added. Synthesis of other T4-coded proteins in the same system continues undisturbed, indicating that the shut-off is specific for gene *32* protein synthesis at this level of added protein. Control experiments showed that the length of the plateau region, as well as the slope of the "shut-off" transition, is independent of the concentrations of the various components of the cell-free translation system. Added dsDNA also does not alter either the length of the plateau or the slope of the shut-off curve. However, added poly(rU) does alter the slope, again without affecting the length of the plateau (see Figs 5 to 7 of Lemaire *et al.* (1978) for further information and details).

These cell-free translation repression experiments were conducted at 37°C and in a complex buffer system, containing as its chief ionic components 30 mM-Tris-acetate (pH 7·2), 10 mM-potassium acetate, 100 mM-$NH_4Cl$ and 13 mM-magnesium acetate. We have carried out control gene *32* protein binding experiments with various nucleic acid lattices in this buffer system, using both nucleic acid hyperchromicity changes and intrinsic protein fluorescence quenching to monitor binding. The results of these measurements, as well as of calculated temperature and salt concentration corrections, suggest that the ionic composition of this cell-free translation buffer system is equivalent to $\sim 0{\cdot}25$ M-NaCl. Thus the autoregulated concentration of free gene *32* protein maintained in a T4-infected *E. coli* cell must be close to that established in the cell-free translation repression experiments of Lemaire *et al.* (1978). We use 2 to 3 $\mu$M in this analysis; other arguments supporting a value of this magnitude are presented in the Discussion.

The following conclusions that are central to our analysis were also derived in (or supported by) the work of Lemaire *et al.* (1978). (1) Gene *32* protein binds preferentially to a specific component of the RNA derived from T4-infected cells. Because shut-off is specific for the synthesis of gene *32* protein, this component must be a portion of the gene *32* mRNA. (2) The abruptness with which shut-off occurs as a function of added gene *32* protein suggests that the shut-off (and the binding of the protein to the gene *32* mRNA that is assumed to be responsible for it) is co-operative in gene *32* protein concentration. (3) Single-stranded DNA effectively binds gene *32* protein more tightly than does the gene *32* mRNA operator site. (4) 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 for (unstructured) poly(rU). (5) Double-stranded DNA, and the other components of the cell-free

† Abbreviations used: ssDNA and ssRNA, single-stranded DNA and RNA; dsDNA, double-stranded DNA.

translation system, bind gene *32* protein less strongly than does the gene *32* mRNA operator. (6) 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 synthesis of other T4 proteins in the cell-free translation system, suggesting that the gene *32* mRNA operator site for gene *32* protein binding probably differs only quantitatively from translational control sites on other T4 mRNAs.

In the next section we use the known binding parameters of gene *32* protein to various nucleic acid squences (as well as the known stabilities of various types of partially and totally double-stranded DNA and RNA lattices), to calculate titration curves for the binding of gene *32* protein to various potential nucleic acid targets under physiological conditions. The results are fully and quantitatively compatible with the experimental facts outlined above.

## 4. The quantitative model

### (a) *Calculation of gene* 32 *protein binding curves* in vivo *for physiological nucleic acid targets*

Nothing we know about gene *32* protein suggests that it might carry an as yet undiscovered (and very tight) binding affinity for some very special single or double-stranded nucleic acid sequence or special element of "tertiary" nucleic acid structure. Thus we proceed on the "unglamorous" (Doherty *et al.*, 1982) basis that it binds preferentially (and co-operatively) to single-stranded regions of nucleic acid lattices, with a net binding affinity *in vivo* that is calculable using equations (1) and (2) and the data of Table 1. The binding parameters summarized in Table 1 suggest qualitatively that the higher values of $K\omega$ for ssDNA sequences relative to ssRNA sequences of the same base composition may account for the saturation of the former sequences at lower free protein concentrations than the latter. However, these data alone do not suggest a molecular basis for the preferential binding of gene *32* protein to its own mRNA, unless perhaps a site on that mRNA is much richer than the average sequence in rG residues (see Table 1). Sequence data on gene *32* mRNA (see below) show that this is *not* the case.

The role of nucleic acid secondary structure must also be considered. Many studies *in vivo* and *in vitro* have shown that single-stranded DNA, and particularly RNA, is highly structured. Thus, under physiological conditions we would expect these lattices to contain many regions of intrachain (hydrogen-bonded and base-paired) secondary structure (i.e. "hairpins"). Such regions of secondary structure (and often of superimposed tertiary structure as well) are not only very prevalent in transfer and ribosomal RNA, they are crucial in the formation of these entities into biologically active structures. More indirect data on mRNA structure and function suggest that these entities, in their functional forms, are also highly structured (see Gold *et al.*, 1981).

The transiently ssDNA sequences formed in DNA replication are also likely to contain an appreciable fraction of hairpin structures; in fact, the current view of gene *32* protein in replication suggests that one of the primary roles of the protein in this process is to "melt-out" these adventitious structures. Thus, clearly the secondary structures of the target nucleic acid lattices are involved as well, and the

relative "strengths" (conformational free energies) of these structures must be "balanced" against the physiologically maintained concentration of gene *32* protein to permit the complete melting-out of DNA hairpins while retaining at least those hairpins of mRNA that are crucial to its biological function.

### (b) *Calculation procedures*

The conformational stability of various duplex nucleic acid structures can be estimated using the approach and free energy parameters developed by Crothers, Tinoco and co-workers (see Materials and Methods). The thermodynamic stability of various elements of nucleic acid secondary structure can then be calculated as a function of free gene *32* protein concentration. Figure 1 outlines the overall models on which our calculations are based.

For each potential nucleic acid binding lattice, we first calculate the conformational free energy ($\Delta G^0_{conf}$) that stabilizes the particular element of secondary structure under consideration. ($\Delta G^0_{conf}=0$ for an initially "open", i.e. single-stranded, sequence.) This establishes the magnitude of the unfavorable (to gene *32* protein binding) free energy that must be overcome by the free energy of complex formation. We then calculate, as a function of free protein concentration, the binding free energy ($\Delta G_{bind}$) associated with the (co-operative) binding of gene *32* protein to all the portions of the structure that are not accessible to the protein ligand in the folded (duplex) form of the polynucleotide lattice.

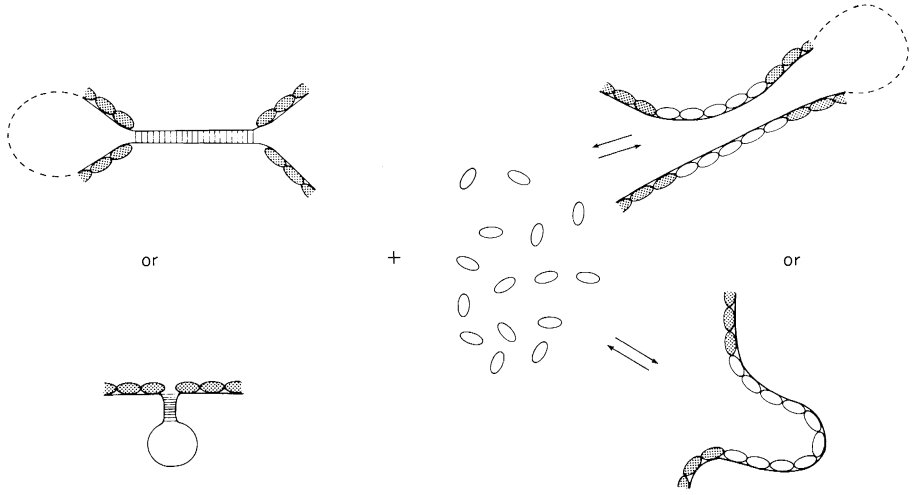


FIG. 1. Model for the 2-state "infinite lattice" calculations. The upper reaction illustrates the melting and complexation (with binding protein) of a stretch of base-pairs of duplex DNA (or RNA) located within a long unpaired (and gene *32* protein-coated) nucleic acid sequence. The lower reaction illustrates the same process for a partially duplex stem-loop structure where neither the initially base-paired "stem" nor the single-stranded "loop" can bind protein in the ordered form (the loop in the ordered structure does not bind protein prior to melting because the looped segment is too short or too conformationally restricted to permit interaction with the protein).

The equilibrium constant (and standard free energy change) associated with the transconformation reaction ($NA_{ds} \rightleftharpoons NA_{ss}$) may be written:

$$K_{conf} = \frac{[NA_{ss}]}{[NA_{ds}]}; \quad \Delta G^0_{conf} = -RT \ln \frac{[NA_{ss}]}{[NA_{ds}]}, \tag{3}$$

where $[NA_{ds}]$ and $[NA_{ss}]$ represent, respectively, the molar concentration of duplex and open (single-stranded) nucleic acid lattice (in units of nucleotide residues). The equilibrium constant of the subsequent binding reaction:

$$NA_{ss} + mP \rightleftharpoons NA_{ss}P_m \tag{4}$$

may be written:

$$K_{bind} = \frac{[NA_{ss}P_m]}{[NA_{ss}][P]^m} \tag{5}$$

where $[NA_{ss}P_m]$ and $[Na_{ss}]$ are the concentrations of lattice sites (in units of nucleotide residues) complexed and uncomplexed at equilibrium, $m$ is the number of protein ligands (of site size $n$) required to cover the segment of polynucleotide lattice exposed in the transconformation reaction, and $[P]$ is the equilibrium free protein concentration (in units of protein monomers). The net binding free energy is then:

$$\Delta G_{bind} = -RT \ln K_{bind} + RT \ln \frac{[NA_{ss}P_m]}{[NA_{ss}][P]^m} \tag{6}$$

and the net free energy change of the coupled unfolding and binding process is:

$$\Delta G_{net} = \Delta G_{bind} - \Delta G^0_{conf} \tag{7}$$

(we set $\Delta G_{conf} = \Delta G^0_{conf}$ by definition, since we are dealing with intramolecular conformational changes). The equilibrium constant for the overall process is:

$$K_{net} = \frac{[NA_{ss}P_m]}{[NA_{ds}][P]^m} = (K_{conf})(K_{bind}) \tag{8}$$

and the fraction of the original duplex structure converted to single-stranded nucleic acid–protein complex is:

$$\theta = \frac{[NA_{ss}P_m]}{[NA_{ds}] + [NA_{ss}P_m]} = \frac{(K_{conf})(K_{bind})[P]^m}{1 + (K_{conf})(K_{bind})[P]^m}. \tag{9}$$

Finally:

$$K_{bind} = (K\omega)_1 (K\omega)_2 \ldots (K\omega)_m = \prod_{i=1}^{i=m} (K\omega)_i. \tag{10}$$

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

Using this procedure, we can calculate the effective stability ($\Delta G_{net}$) of any specific base-paired (or partially base-paired) structure as a function of the concentration of free binding protein, $[P]$, or alternatively, we can calculate the free protein concentration required to overcome the stability of a given nucleic acid structure.

The values of $K\omega$ that apply under physiological conditions and have been used in these calculations are collected in Table 1†.

(c) *Melting and complexation of fully duplex nucleic acid structures*

Figures 2 and 3 contain calculated (two-state) model binding curves for the melting and complexation with gene *32* protein of fully duplex (containing no single-stranded loops not complexed with protein) DNA and mRNA structures such as that illustrated in the upper portion of Figure 1. For most of the calculations, we melt segments 21 base-pairs long; i.e. $m=6$ (additional) proteins bound to the open form of these structures. The binding curves marked dsT4DNA and dsT4mRNA in Figures 2 and 3 were calculated using an arbitrary repeated sequence that approximates the average base composition of T4 DNA, and the corresponding average mRNA (see Figure legends). The results show that melting is quite abrupt (due to the "co-operative filling-in" of the open lattice segments on melting); the degree of sharpness of the saturation of the lattice in terms of free protein concentration ($[P]_{\text{free}}$) depends only (in the two-state model) on the number of proteins ($m$) involved in the reaction. In each case, the DNA structure undergoes equilibrium melting at lower free protein concentrations than does the homologous mRNA segment; this is a consequence of the fact that $K\omega$ is larger for deoxyribopolynucleotides than for ribopolynucleotides (Table 1). Duplexes containing dG·dC (or rG·rC) base-pairs melt at significantly higher free gene *32* protein concentrations than do duplexes containing dA·dT (or rA·rU) base-pairs. This difference is due to the much greater intrinsic stability toward melting of G·C-containing duplex structures, and not to differences in $K\omega$ for these sequences; Table 1 shows that the average value of $K\omega$ for a G plus a C residue is about equal to that for an A plus a T (or U) residue. We calculate (results not shown) that the autoregulated concentration of gene *32* protein is too low to melt *any* fully duplex DNA or mRNA structure, including those containing A·T or A·U base-pairs only.

(d) *Melting and complexation of partially duplex DNA structures*

In Figure 2 we also show the degree of melting (and thus of saturation) of various initially unsaturated and partially structured T4 DNA sequences of average base

† In all these calculations, we assume that the originally base-paired and looped (or bulged) nucleic acid structures are fully unfolded in the first step, and that in this transconformation process they pass from being fully inaccessible to being fully accessible to binding proteins. For convenience, we also assume that these structures occur within long nucleic acid lattices, which are otherwise fully complexed with binding protein, to avoid end effects and to let each protein that binds to the lattice contribute a full factor of $K\omega$ to $K_{\text{overall}}$ (this assumption will be relaxed for finite lattices, below). Furthermore, in some cases $m$ (the number of proteins bound in the second step) is not an integer, meaning that the total number of bases ($N$) involved in the initial non-liganded nucleic acid structure may not be equal to an integral number of protein units of length $n$. In this case, we simply use the appropriate fraction of $K\omega$, and assume that the ligands shift to accommodate the extra fractional protein elsewhere on the lattice. Also, we assume that the nucleic acid stability parameters, which were originally measured largely using oligoribonucleotides of known structure and sequence, apply equally well to polyribo- and polydeoxyribonucleotides. Finally, we note that the thermodynamic approach taken here is, by definition, a 2-state model, with all duplex (or partially duplex) structures in the system assumed to be either fully in the initial (unmelted) state or fully complexed with binding protein. It turns out that these assumptions are very appropriate for the calculations conducted here; i.e. the abrupt transition (with increasing free gene *32* protein) to saturation of short duplex or hairpin regions internal to a relatively long and previously saturated polynucleotide (Fig. 1).

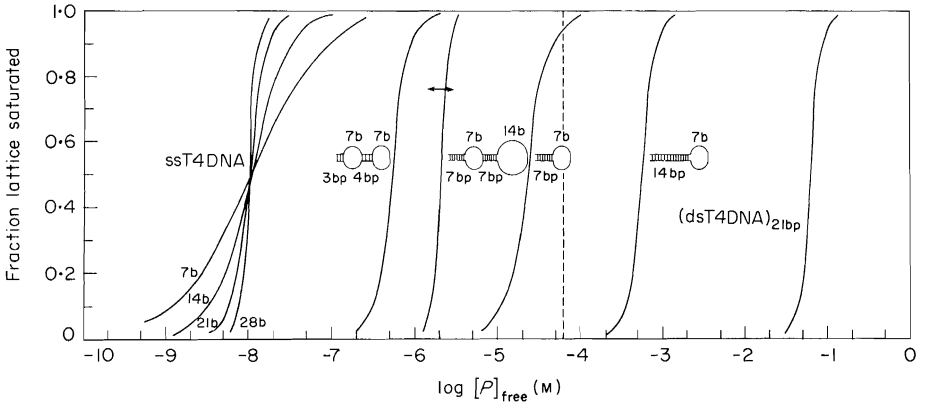


FIG. 2. Calculated 2-state binding curves for the "melting" and complexation by gene *32* protein of various "looped" and "bulged" T4 DNA structures, plotted as a function of free gene *32* protein concentration. The double-ended arrow indicates the estimated concentration range of free intracellular gene *32* protein *in vivo* (see the text). The titration curves correspond, respectively, to the indicated stem-loop (and/or bulge) structures. The DNA used in these calculations consists of tandem repeats of an arbitrary sequence (A-C-G-G-T-A-A) of average T4 DNA base composition. The titration on the left shows the "sharpening effect", in the 2-state model, of increasing the length of DNA lattice. (A complete infinite lattice binding calculation, including both overlap of potential binding sites and binding co-operativity (McGhee & von Hippel, 1974), generates a binding isotherm that is essentially superimposable on the 28b line for ssT4 DNA.) The broken vertical line at $\sim 8 \times 10^{-5}$ M-free protein indicates the approximate "cut-off" concentration at which gene *32* protein would begin to bind appreciably to the exterior of double-stranded T4 DNA (see the text). The curve on the right, labelled dsT4DNA, corresponds to the binding isotherm for protein binding to ssDNA formed by melting the initially fully dsDNA structure. b, Bases; bp, base-pairs.

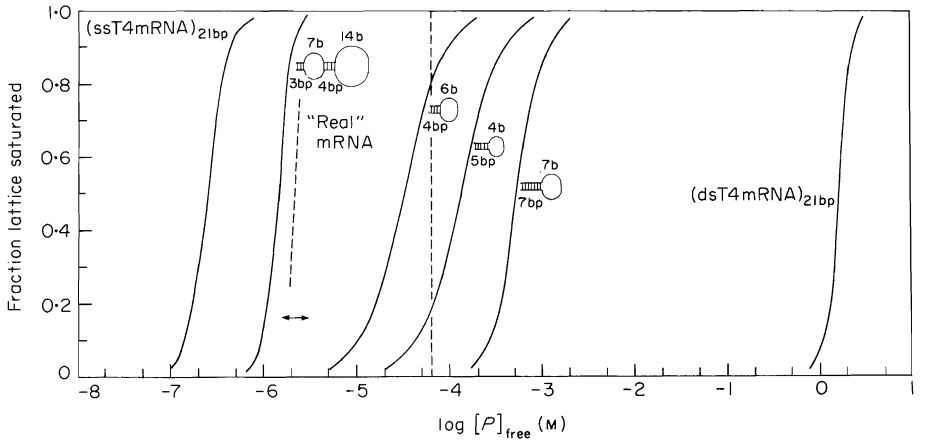


FIG. 3. Binding curves for the "melting" and complexation by gene *32* protein of various hypothetical initially looped and bulged T4 mRNA structures. Symbols and other details are as for Fig. 2. The sloped broken line labelled "Real" mRNA is the approximate binding isotherm for the gene *32* mRNA operator site, as estimated from the Lemaire *et al.* (1978) experiments (see the text). b, Bases; bp, base-pairs.

composition. The results show that all accessible single-stranded sequences (marked ssT4DNA) are fully saturated under intracellular conditions at free gene *32* protein concentrations greater than $\sim 0{\cdot}05\ \mu$M. Figure 2 shows that most secondary structures that might form as adventitious stem-loop (hairpin) structures in ssDNA sequences transiently exposed in the course of replication, recombination or repair will have been melted to completion and saturated with gene *32* protein under intracellular conditions at free protein concentrations of $\sim 2$ to $3\ \mu$M. Figure 2 suggests that DNA hairpins containing very long stems or large stem-to-loop ratios may be stable at the autoregulated gene *32* protein concentration; we discuss below the possible occurrence of such very stable DNA hairpins in the replication process. We note, as expected for the two-state model, that hairpins containing more (total) residues melt with increased apparent co-operativity.

### (e) *Melting and complexation of partially duplex mRNA structures*

Figure 3 shows the results of similar calculations for partially duplex elements of secondary structure for T4 mRNA of average composition. We see, as a consequence of the weaker binding of gene *32* protein to RNA sequences (Table 1), that virtually all hairpins (with the exception of those containing more than $\sim 70\%$ single-stranded residues) are stable to melting (by gene *32* protein) under physiological conditions. This fact makes it possible to consider models of mRNA, differing only in the extent and placement of secondary structure, which can, in principle, be discriminated by co-operative single-stranded nucleic acid binding proteins.

The first phase of the gene *32* protein autoregulatory cycle requires that ssDNA sequences (and hairpin loops) be complexed to completion prior to protein binding to RNA. This condition is fully met as a consequence of the tighter affinity of gene *32* protein for DNA, together with binding co-operativity (compare the single-stranded DNA and RNA binding curves of Figs 2 and 3). We next consider the possible nature of the target site on gene *32* mRNA that results in its effective binding saturation (and translational shut-off) at a gene *32* protein concentration lower than that which shuts off the translation of the other T4 mRNAs. The original model for the autoregulation of gene *32* protein synthesis (Russel *et al.*, 1976) proposed that the operator might overlap the gene *32* mRNA ribosome binding site, and that this operator site might be largely unstructured. Such operator sites could, of course, comprise (or be located within) specific hairpins, but in view of the known preference of gene *32* protein (and probably of the ribosome as well) for single-stranded sequences it seemed more likely that the critical site (or sites) on gene *32* mRNA should be a largely unstructured sequence (see also Lemaire *et al.*, 1978; Newport *et al.*, 1981*a*,*b*).

How might this site (the translational operator) be functionally discriminated from other single-stranded sequences on this mRNA or others?† In principle, there are two possibilities: either the gene *32* protein mRNA control sequence has an average composition that leads to unusually tight binding (Table 1 suggests that this

† Functional discrimination here means that this site binds gene *32* protein to completion, and thus shuts-off further synthesis, at concentrations of free gene *32* protein too low to complex the binding sites on other T4 mRNAs.

might be best accomplished with a very G-rich sequence), or the control sequence must consist of a particularly long single-stranded region unencumbered by hairpins that are stable at physiological concentrations of gene *32* protein.

The reason that the length of such a single-stranded sequence is important is that if this sequence is bounded by stable hairpins at both ends, or equivalently by a stable hairpin at one end and a chain end at the other, then we are dealing with binding to a finite lattice, which differs importantly from the effectively infinite nucleic acid lattices that we have been considering to this point.

### (f) *Finite lattice binding*

In binding to an effectively infinite lattice, each gene *32* protein molecule contributes one "unit" of $K\omega$ to the total value of $K_{\text{bind}}$ (eqn (10)). For a finite lattice, on the other hand, we may write:

$$K_{\text{bind}} = K_1(K\omega)_2(K\omega)_3 \ldots (K\omega)_m = K_1 \prod_{i=2}^{i=m} (K\omega)_i \tag{11}$$

or, for a finite lattice of constant composition:

$$K_{\text{bind}} = K(K\omega)^{m-1}. \tag{12}$$

The consequence of the loss of the factor of $\omega$ for the first (or last) protein that binds to the lattice is shown in Figure 4; clearly, as the length of the lattice decreases binding becomes less and less co-operative, and also more free gene *32* protein is required to saturate the site†. Thus a uniquely long single-stranded finite lattice region would saturate first, and become an excellent candidate for the gene *32* mRNA operator.

As originally predicted (Russel *et al.*, 1976), and as we will show below, a sequence of exactly such properties is, in fact, found overlapping the initiation codon of T4 gene *32* protein mRNA (Krisch *et al.*, 1980; Krisch & Allet, 1982). Thus Figure 4 suggests that the autoregulatory target site could be a completely unstructured single-stranded DNA sequence $\sim$30 nucleotide residues ($m \simeq 4$) in length and of average base composition (in terms of $K\omega$ values). This sequence would be flanked by stable hairpins (see Fig. 4, inset). Alternatively, the operator could be a somewhat longer sequence containing some weak secondary structure that would be "melted out" at the autoregulated free gene *32* protein concentration. We suggest that potential operator sequences of the other T4 mRNA messages consist of shorter (and/or more structured) single-stranded sequences, and thus remain essentially uncomplexed at the regulated intracellular gene *32* protein concentration.

† The 2-state approach used throughout this paper is quantitatively less accurate for finite than for infinite lattice calculations, though the important qualitative results are clearly reflected in such calculations (broken curves in Fig. 4). The 2-state results are somewhat incorrect, because initially "open" finite lattices will saturate in discrete stages; thus at half saturation the 2-state model postulates that one-half of the lattices are totally saturated, and that the other half are totally "empty". Actually, the lattices contain a distribution of different levels of saturation under these conditions, and this element is lost in the 2-state calculations. Statistical mechanical calculations (unbroken curves, Fig. 4) that take this distribution into account (Epstein, 1978), show that the binding curves for finite lattices that are calculated by the 2-state model in Fig. 4 are somewhat shifted and altered in shape relative to the "exact" transitions.

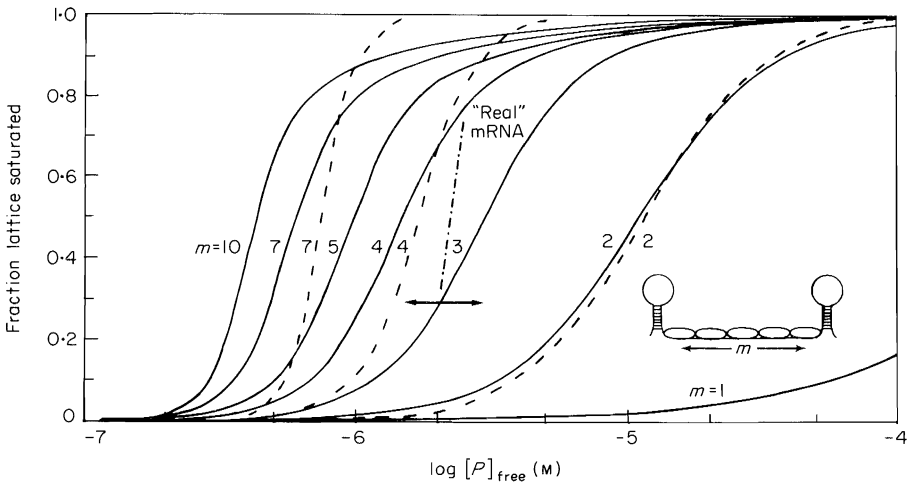


FIG. 4. Binding curves for the finite mRNA lattices of varying length. The broken curves represent the 2-state approximation, calculated as outlined in the text. The unbroken curves were calculated by the "exact" method of Epstein (1978); for further details, see Newport *et al.* (1981*b*) and the text. The lengths of the lattices are defined in units ($m$) of protein monomer binding sites. (The site size of gene *32* protein binding co-operatively in the polynucleotide binding mode is 7 nucleotide residues. Thus the lengths of the respective finite lattices, in units of nucleotide residues, are $7m$.)

### (g) *Double-stranded DNA binding*

As noted above, gene *32* protein also can bind (non-co-operatively) to the "exterior" of dsDNA (Jensen *et al.*, 1976). An estimated titration curve for such binding is also indicated in Figures 2 and 3. Clearly, dsDNA does not bind gene *32* protein at or below the autoregulated free protein concentration; however, an appreciable "overshoot" in this concentration could result in binding to the unencapsulated dsDNA free in the infected *E. coli* cell. Such an overshoot might occur in regulatory mutants in which the gene *32* mRNA operator sequence is partially deleted or "overstructured", etc. We have no experimental evidence that suggests the actual existence of such mutants at present, but such binding to dsDNA could serve as a secondary mechanism to limit the concentration of free gene *32* protein in the event of partial failure of the primary control system.

## 5. Comparison with sequence data

### (a) *Identification of the gene* 32 *mRNA operator sequence*

The recent determination (Krisch *et al.*, 1980; Krish & Allet, 1982) of the T4 DNA sequence coding for gene *32* protein makes it possible to test further the validity of the "unstructured operator" hypothesis. The sequence surrounding the initiation codon of the gene *32* message is shown in Figure 5. In most mRNA sequences this region contains the ribosomal binding site at which mRNA translation is initiated (see Gold *et al.*, 1981), and thus comprises the most logical candidate for the gene *32* mRNA operator site. This view is based on the simplest repression model in which

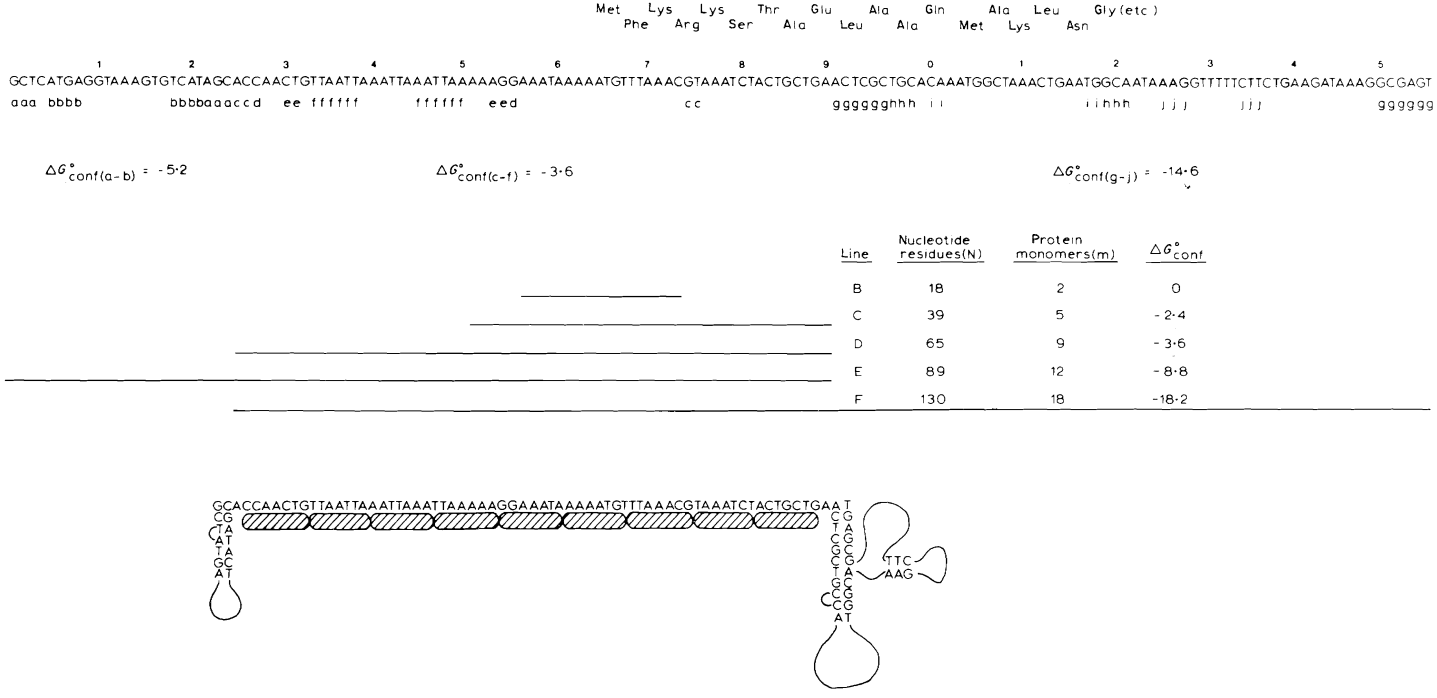


FIG. 5. Sequence and conformational stability of the putative gene *32* mRNA operator site and vicinity. At the top we show the DNA sequence (non-coding strand only; the sequence as written corresponds to mRNA when T is replaced by U), with the beginning of the gene *32* protein sequence written above. 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 values of $\Delta G^{0}_{conf}$ below correspond to the calculated stability of the indicated hairpin structures. The lines (labelled B through F) correspond to the segments tested as potential operator sites (see the 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. The sequence hyphens have been omitted for clarity.

gene *32* protein (as repressor) and the ribosome (and/or ribosome accessory proteins) bind competitively to this operator-initiator site.

The sequence of gene *32* mRNA in the vicinity of the initiation codon is remarkable, even for a phage carrying 66% adenine plus thymine residues. As Figure 5 shows, the ribosome binding site region contains a stretch of 40 nucleotides (residues 33 to 72 inclusive) in which the only nucleotides other than A or U are the three nearly essential G residues that participate in the Shine–Dalgarno sequence and in the initiation codon (see Gold *et al.*, 1981). We have computed $\Delta G^0_{\text{conf}}$ for a variety of arbitrary segments within the gene *32* initiation sequence, in order to determine whether an unstructured domain could exist in this region that is of sufficient length to serve as an operator site within the quantitative constraints imposed in this paper. A related goal, of course, is to ask whether such a domain (if found) is unusual (or even unique) among the available sequences representing other regions of the T4 genome.

Values of $\Delta G^0_{\text{conf}}$ for various mRNA segments have been estimated using a secondary structure calculation algorithm (see Materials and Methods). Some results are presented in Figure 5. We note that calculations such as these will be incorrect in detail and may, in some cases, be grossly incorrect (see further discussion in the next section). The calculated values will change as nucleotide residues are added or subtracted from either end of a particular lattice segment, and changes will occur that alter even the specific sets of nucleotides thought to be involved in base-pairing. Nevertheless, no calculation that we have carried out ever places the gene *32* initiation codon into a secondary structure (see below). Thus we proceed to calculate the free gene *32* protein concentration that would be reached (i.e. at which the system would autoregulate) if the operator comprised sequences of various lengths in the putative gene *32* mRNA ribosomal binding region (Fig. 5).

The amount of gene *32* protein in a T4-infected cell is not sufficient to titrate more than a small fraction of the intracellular T4 mRNA (Gold *et al.*, 1977). Therefore we use the finite lattice approach to calculate the expected values of $[P]_{\text{free}}$ for various trial operators. For this situation:

$$\Delta G_{\text{bind}} = -RT[m \ln(K\omega) - \ln \omega]. \tag{13}$$

We define:

$$\rho = \frac{[\text{NA}_{\text{ss}}P_m]}{[\text{NA}_{\text{ds}}]}, \tag{14}$$

so that:

$$K_{\text{net}} = \rho[P]^{-m}$$

and

$$\Delta G_{\text{net}} = -RT \ln(\rho[P]^{-m}). \tag{15}$$

We then combine equations (7), (13) and (15) to obtain:

$$\Delta G^0_{\text{conf}} = -RT(m \ln(K\omega) - \ln \omega + m \ln[P] - \ln \rho). \tag{16}$$

Inserting the value of $K\omega$ that applies to average T4 mRNA (as defined in the legends to Figs 2 and 3) under physiological conditions, we partially solve equation (16) to obtain:

$$\Delta G^0_{\text{conf}} = 0{\cdot}62\{7{\cdot}60 + \ln \rho - m(\ln[P] + 15{\cdot}2)\}. \quad (17)$$

We next assume, for convenience, that $\rho$ is $\simeq 10$ (equivalent to $\theta \simeq 0{\cdot}9$) when gene *32* protein synthesis is effectively repressed (i.e. we define the autoregulated value of $[P]_{\text{free}}$ as being that which is attained when $\rho \simeq 10$)†. The accuracy of our approach is limited by the fact that, for simplicity, we use the two-state finite lattice approximation here. This also introduces small errors in the calculated values of $[P]_{\text{free}}$ that apply at $\rho = 10$, depending on the magnitude of the difference between the two-state and the exact finite lattice titration curve. The magnitude of this difference will depend on the size ($m$) of the finite lattice segment (see Fig. 4).

The longest totally unstructured gene *32* mRNA domain extends from nucleotide 56 to 73, inclusive (line B, Fig. 5), and thus corresponds to an $m$ value of $\sim 2{\cdot}5$. (The site size, $n$, for gene *32* protein is $\sim 7$ nucleotide residues.) Using equation (17), we calculate that $[P]_{\text{free}}$ would have to reach $\sim 35\ \mu$M to saturate (to $\rho = 10$) this sequence. Just 5′ and 3′ to this minimal, totally unstructured, putative operator are regions of only marginal structural stability; for this sequence (line C, Fig. 5) $\Delta G^0_{\text{conf}} = -2{\cdot}4$ kcal/mol (at $m = 5$), and a value of $[P]_{\text{free}}$ of only $\sim 4\ \mu$M would be required to saturate this sequence (at $\rho = 10$). This calculation graphically illustrates the finite lattice effect; a longer, partially structured lattice is saturated at a lower free gene *32* protein concentration than is required to fill the minimal site, in spite of the presence of some secondary structure that must be overcome in titrating the larger site.

The proposed operator sequence can be extended still further in the 5′ direction (line D, Fig. 5); for $m = 9$ and $\Delta G^0_{\text{conf}} = -3{\cdot}6$ kcal/mol, $[P]_{\text{free}} \simeq 1{\cdot}5\ \mu$M. No further extension of the operator (beyond line D) in the 5′ direction seems likely, because a stable hairpin (in line E, Fig. 5) prevents the filling of a longer sequence. $\Delta G^0_{\text{conf}}$ for this hairpin alone is $-5{\cdot}2$ kcal/mol, and thus would require appreciably higher values of $[P]_{\text{free}}$ to melt-out. Further extension of the postulated operator site in the 3′ direction (beyond line D, Fig. 5) also seems unlikely, since this direction is also closed off by a very stable hairpin (see line F, Fig. 5). We thus conclude that the optimal sequence for gene *32* protein binding, and thus the most likely candidate for the gene *32* mRNA operator site, corresponds to line D of Figure 5.

This putative operator region is remarkable not only for its richness in A and T, but also because it appears to be uniquely unstructured. It includes the presumed Shine–Dalgarno sequence and the first nine triplets that encode the protein, and thus is a logical candidate for the ribosome binding site for gene *32* protein translation, in keeping with the simplest (direct competition between gene *32* protein and ribosome binding) repression model outlined above. This operator region, showing the terminating hairpins and the central region coated with nine

† This assumption is justified at the degree of precision at which we are operating here, since eqn (17) is not very sensitive to small changes in $\rho$ and, in addition, many experiments *in vivo* show that derepression of gene *32* protein translation is $\sim 10$-fold when new single-stranded DNA sequences are made available (Russel *et al.*, 1976).

gene *32* protein molecules, is presented in schematic form at the bottom of Figure 5. This model is fully consistent with the very recent results presented by Krisch & Allet (1982), who have shown that DNA deletions that remove all but the most 5′ portion of the gene *32* protein coding sequence itself do not prevent repression of the translation of the remaining fragment by active gene *32* protein in a cell-free translation system. The notion that the operator region contains some structure that must be melted-out on binding is also compatible with the observation (cited by Lemaire *et al.* (1978)) that repression of translation by gene *32* protein *in vivo* is essentially temperature independent. If the operator were totally unstructured, the decrease in binding affinity of gene *32* protein with increasing temperature (see above) would result in a decreased apparent value of $[P]_{free}$ required for repression at the lower temperatures examined.

### (b) *Is the gene* 32 *mRNA operator sequence unique?*

We next asked quantitatively whether this proposed operator sequence is unique among T4 sequences in its lack of secondary structure, and thus in its effectively increased affinity for gene *32* protein. To this end, we have used the entire catalogue of available T4 nucleic acid sequences (approximately 5% of the entire genome; see Materials and Methods), and calculated $\Delta G^0_{conf}$ for a "rolling window" (i.e. a "moving" finite lattice) of lengths 30, 40 and 50 nucleotide residues. The results are presented in Figure 6, as a plot of $\Delta G^0_{conf}$ *versus* the fraction of sequences with $\Delta G^0_{conf}$ that is smaller (more negative; i.e. corresponding to more structure) than the indicated value. Clearly, long unstructured domains are not very common in the T4

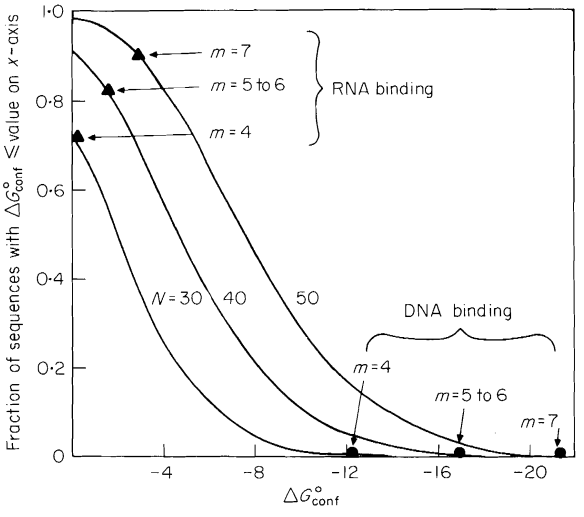


FIG. 6. Plot of $\Delta G^0_{conf}$ for various T4 DNA (or RNA) sequences as a function of the fraction of total sequences that contain more structure (i.e. that are characterized by a more negative value of $\Delta G^0_{conf}$) than the indicated values. The 3 curves are calculated for lattice "windows" that are 30, 40 and 50 nucleotide residues in length, respectively. The points estimate the minimum fractions of the total lattice segments (as DNA or RNA) that *cannot* bind gene *32* protein (at the indicated values of $m$) at the autoregulated gene *32* protein concentration and intracellular conditions (see the text).

genome. Less than 2% of the genomic sequences are unstructured at lattice lengths of 50 residues; the predicted operator (line D, Fig. 5) has much less structure than other sequences of comparable length. We have also looked specifically at more than ten T4 ribosome binding site sequences; none is as unstructured as the proposed gene *32* mRNA operator.

We note, however, that more than 25% of the T4 sequences appear to be unstructured at a lattice length of 30 nucleotide residues (see Fig. 6). Clearly, the 30-residue calculation underestimates the degree of secondary structure in RNA. For example, the hairpin that closes the gene *32* operator on the 3′ side of the initiation codon is completely missed in the 30 (and the 40 and 50) residue analysis; in fact, only when about 65 residues are analysed in our program does the entire stable structure, including lattice-terminating hairpins, appear in the calculation (Fig. 5). We note that fairly complete secondary structure information now exists for some *E. coli* 16 S rRNA molecules (Noller, 1980); here also one would underestimate secondary structure by sequential analysis of rRNA domains less than 50 residues in length.

We have used the results of Figure 6 to calculate the fraction of the RNA sequences to which gene *32* protein could bind (to a fractional saturation ($\theta$) of $\sim 0{\cdot}5$ or more) at a value of $[P]_{\text{free}}$ of $\sim 2$ $\mu$M under physiological salt and temperature conditions. The results indicate that only when $m \geqslant 5$ can gene *32* protein overcome any RNA secondary structure at all (see also Fig. 4). The fraction of RNA sequences (at lattice lengths of 30, 40 and 50 residues) that cannot be saturated with gene *32* protein at the autoregulated concentration and intracellular conditions are also indicated in Figure 6. These values range from 72% to 92% of the total putative T4 mRNA sequences; we note that these numbers represent an appreciable overestimate because of the short lattice lengths used in the calculation (see above).

The amount of T4 mRNA present in an infected *E. coli* cell was calculated by Gold *et al.* (1977). When one subtracts the mRNA nucleotides that are covered with ribosomes, we estimate that the total amount of mRNA available to gene *32* protein per cell is approximately $1{\cdot}3 \times 10^6$ nucleotide residues. If we assume that only 1% of this mRNA is "operator-like" (i.e. complexed by gene *32* protein at the free protein concentration at which the "true" operator is saturated), $\sim 2000$ molecules of gene *32* protein will be complexed by the total T4 mRNA. Using $\sim 10^{-15}$ liters as the volume of an *E. coli* cell (von Hippel *et al.*, 1974), the infected cell would contain $\sim 3$ $\mu$M-gene *32* protein bound to mRNA. The suggestion that non-initiation regions of other T4 mRNAs might serve as an immediately available reservoir for mobilization of gene *32* protein when needed to bind to newly formed ssDNA sequences (Gold *et al.*, 1977) thus remains a viable possibility.

### (c) *Can all (or most) ssDNA sequences be saturated at the autoregulated gene* 32 *protein concentration?*

Calculations shown in Figure 6 also suggest that virtually all the ssDNA domains that appear during replication or other physiological processes can be complexed by gene *32* protein. Both because ssDNA is presumed to complex with gene *32* protein in the infinite lattice mode as the replication fork moves (i.e. the new ssDNA exposed in

the "rolling" replication "window" is probably flanked by previously bound gene *32* protein or other T4 replication proteins with which gene *32* protein can interact co-operatively), and because the binding constant for ssDNA is higher than that for ssRNA (Table 1), very stable DNA hairpin structures are expected to denature at $\sim 2$ $\mu$M-free gene *32* protein. We have indicated in Figure 6 the fraction of DNA sequences (of lengths 30, 40 or 50 residues) that would be expected to survive melting (by gene *32* protein binding) at this level of free protein. These levels are very close to zero for all calculated lattice lengths, if we assume that the total available gene *32* protein exceeds the total amount of intracellular ssDNA present. Measurements of the total concentrations of intracellular ssDNA and gene *32* protein have been made (Gold *et al.*, 1977); gene *32* protein is in excess and thus the assumptions of the infinite lattice calculation are, at least in this respect, legitimate.

## 6. Discussion

### (a) *General principles*

The calculated titration curves of Figures 2, 3 and 4, in which we have plotted the expected fractional saturation of various structured, unstructured and partially structured polynucleotide lattices with gene *32* protein under intracellular conditions, pass from the essentially free to fully complexed state over a relatively narrow range of free protein concentration. As shown in Table 1, the differences in intrinsic protein–nucleic acid binding affinity for the various polynucleotide lattices are not large; however, due to binding co-operativity, the transitions between the free and the saturated state are quite abrupt, and thus the transitions for the various lattices are effectively separated along the free protein concentration axis. As a consequence, an "on-off switch", based on the saturation of a particular sequence (here the gene *32* mRNA translational operator site) can effectively permit the total saturation of lattices that bind to completion at lower free protein concentrations, while leaving lattices that saturate at higher concentrations totally unencumbered. The position of a particular transition along the free protein concentration axis (Figs 2 and 3) depends only on the intrinsic binding affinity of the lattice segment ($K\omega$) and on the amount of conformational free energy (as secondary structure) that must be overcome to transform the segment to a fully open state suitable for gene *32* protein binding. The position of the center of the transition does not (to a first approximation) depend on the length of the segment to be saturated if this segment is located within a longer lattice that has previously been saturated with gene *32* protein at a lower free protein concentration. This corresponds to "infinite lattice binding" conditions, meaning (for a lattice segment of constant composition that is $m$ protein units in length) that $K_{\text{bind}} = (K\omega)^m$, see equation (10).

Both the position of the titration curve along the free protein concentration axis, and the sharpness of the transition itself, will be a function of the length of the lattice segment to be saturated if binding occurs under "finite lattice" conditions (Fig. 4). This means, for a finite lattice of constant composition that is $m$ protein units long, that $K_{\text{bind}} = K(K\omega)^{m-1}$, see equation (12). (A finite lattice is defined as a polynucleotide sequence that is isolated by stable hairpins, and/or by ends of the polynucleotide chain, from other lattice segments to which the protein can bind co-operatively.)

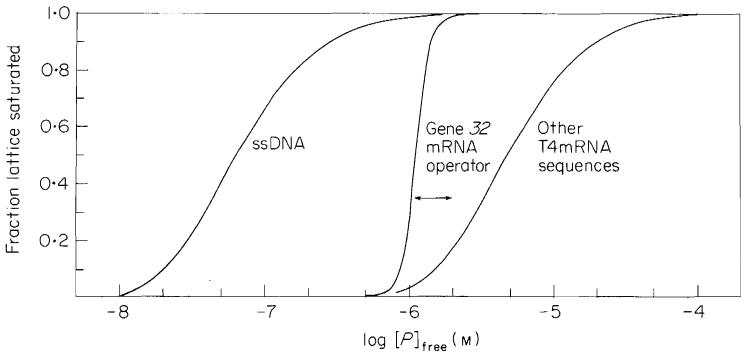


FIG. 7. Binding curves summarizing the gene *32* protein autoregulatory system. The "ssDNA" curve is calculated using the real T4 DNA sequences with an $N = 50$ residue lattice length replication window and the infinite lattice calculation mode. The "gene *32* mRNA operator" curve is calculated for the putative operator structure (line D) shown at the bottom of Fig. 5. The "other mRNA" curve is calculated using real T4 sequences with an $N = 50$ residue lattice length and the finite lattice calculation mode. The other T4 ribosome binding sites that have been examined (Gold *et al.*, 1981; Stormo *et al.*, 1982) are more highly structured, and so should be repressed only at higher concentrations of gene *32* protein, as is observed.

Figure 7 summarizes our calculated results for the functioning of the actual gene *32* protein autoregulatory system in T4 infection. We note that the titration curves for the ssDNA segments and for the "other" (non-gene *32* operator) potential mRNA binding sites are relatively broad, reflecting the real secondary structure heterogeneity of the sequences. In contrast, the titration curve for the proposed gene *32* mRNA operator site itself (line D of Fig. 5) is quite sharp, and binding goes to completion, as it must, at a free gene *32* protein concentration just below the autoregulated level†.

Clearly, the variables considered in this approach provide ample opportunity to "space out" the relevant potential binding targets as a function of free binding protein concentration, and give ample scope for the operation of an autogeneous genome regulatory mechanism of this type, without requiring special affinity of the protein for particular nucleotide sequences. This type of behavior is general, and can apply to any appropriate control system in which free protein concentration is regulated by the sequential binding of the protein to a series of target sequences of decreasing net binding affinity.

### (b) *Applications to other systems*

Gold *et al.* (1981) have described in detail several systems in which protein synthesis appears to be regulated at the translational level, resulting in (usually reversible) negative feedback systems designed to control rather sharply the concentration of the free protein (or proteins) involved. Examples include regulation of several early T4 genes by the *reg*A gene product (Karam *et al.*, 1981), repression of

† The suggestion contained in the positioning of the "other" T4 mRNA binding site titration curve that as much as 8% of the non-gene *32* operator mRNA may be titrated at the autoregulated protein concentration again reflects the "short lattice" artifact (see Fig. 6); i.e. the degree of secondary structure of these sequences is underestimated as described above.

R17/MS2 replicase by the phage coat protein (Spahr *et al.*, 1969), repression of the Qβ coat cistron by Qβ replicase (Weber *et al.*, 1972), and regulation of the expression of the *E. coli* ribosomal proteins (Nomura *et al.*, 1981,1982). It appears obvious to us that the principles defined in this paper should, perhaps in modified form, be applicable to all these regulatory systems. We look forward to the measurement of thermodynamic parameters that will permit the quantitative modelling of some of these systems.

One difference between the gene *32* protein autoregulatory system and some of the others mentioned above (and described in detail by Gold *et al.*, 1981) is that structured, rather than unstructured, RNA sequences may comprise the sequential binding targets. Thus Nomura *et al.* (1981) have proposed that certain ribosomal proteins recognize and bind to particular hairpin structures on the ribosomal rRNA framework in ribosome assembly, and then, presumably at somewhat higher free protein concentrations, recognize and bind to homologous hairpin structures on the relevant mRNA. In this model, the latter binding then acts to repress the further synthesis of a whole set of polycistronically regulated ribosomal proteins.

The various ribosomal proteins bind their rRNA (and probably their mRNA) targets in single copies; how then do we obtain the required "sharpening" of the protein binding curves provided by co-operativity of binding for gene *32* protein to the unstructured operator? (Non-co-operative binding would require a rather large difference in binding affinity to the rRNA and mRNA targets in order to achieve effective regulation, and thus would require a very large free ribosomal protein concentration. This is not observed.) We suggest that one possible solution is that the non-co-operative binding affinities of the individual proteins for their respective rRNA and mRNA targets are about equal, and that it is the "hetero-protein" co-operativity of binding of the other ribosomal proteins to the ribosomal RNA framework that results in the prior saturation of the rRNA target. If correct, this notion would make it possible to carry out titration studies with the ribosomal protein assembly system to define thermodynamically co-operative clusters of ribosomal proteins, which could be compared with the results of other measures of protein distribution in ribosome assembly maps.

### (c) *The autoregulated concentration of free gene* 32 *protein* in vivo

In preceding sections we suggested that gene *32* protein is autoregulated at a free concentration of 2 to 3 μM. This value was estimated initially by extrapolating the repression measurements made *in vitro* by Lemaire *et al.* (1978) to physiological salt concentrations (see above). Two other lines of evidence support this estimate.

Cells infected with gene *46*$^-$ mutants contain no detectable ssDNA; thus in these cells there is no ssDNA "sink" for gene *32* protein. Several experiments (see Russel *et al.*, 1976) suggest that these cells contain a total of ~1000 to 2000 gene *32* protein molecules; this corresponds to ~3 to 4 μM-total protein, which should be mostly free, though some may be bound to non-operator mRNA sequences (see above) or to other T4 replication proteins.

Finally, of course, our initial estimate of $[P]_{free}$ is bolstered by our finding that this concentration of free gene *32* protein falls close to the critical value of $[P]_{free}$ that, under physiological conditions, will melt most DNA hairpins that are expected to

form during replication fork movement, but will not melt most of the secondary structure of the various T4 mRNAs that we assume is required for the function of these entities.

(d) *Why must the free concentration of gene* 32 *protein be regulated?*

Finally, we might ask explicitly, in light of these findings, why the free concentration of gene *32* protein should be autoregulated. A simple answer is that even a modest overproduction of gene *32* protein will shut-off the synthesis of other T4 proteins that are required to be produced in parallel with gene *32* protein in the course of T4 infection. This has been demonstrated in cell-free translation experiments by Lemaire *et al.* (1978). Presumably this occurs because, at higher concentrations of free gene *32* protein, this protein can bind to (and/or melt) mRNA sequences that are either too short or too structured to be complexed at the regulated value of $[P]_{\text{free}}$. In this connection, we note that (unlike many other genome regulatory proteins) no one has succeeded in cloning (and then overexpressing) gene *32* in a living bacterial cell. This could reflect the fact that gene *32* protein will complex and/or melt ssRNA or ssDNA sequences crucial for bacterial function, even at the regulated free protein concentration appropriate for lytic infection by T4 phage.

This work was supported in part by United States Public Health Service research grants GM-15792 and GM-29158 (to PHvH) and GM-19963 (to LG). JWN and LSP were predoctoral trainees of U.S. Public Health Service research training grants GM-07759 and GM-70015. We are very grateful to Dr Henry Krisch and his colleagues for making the DNA sequence of the gene *32* region of T4 available to us prior to publication. We also thank Mr Daniel W. Noble for his help in determining some of the values of $K\omega$ cited here.

## REFERENCES

deHaseth, P. L., Lohman, T. M. & Record, M. T. Jr (1977). *Biochemistry*, **16**, 4783–4790.

Doherty, D. H., Gauss, P. & Gold, L. (1982). In *Multi-Functional Proteins* (Kane, J. F., ed.), CRC Press, Cleveland. In the press.

Epstein, I. R. (1978). *Biophys. Chem.* **8**, 327–339.

Gold, L., O'Farrell, P. A. & Russel, M. L. (1976). *J. Biol. Chem.* **251**, 7251–7262.

Gold, L., Lemaire, G., Martin, C., Morrisett, H., O'Connor, P., O'Farrell, P. Z., Russel, M. & Shapiro, R. (1977). In *Nucleic Acid-Protein Recognition* (H. J. Vogel, ed.), pp. 91–113, Academic Press, New York.

Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B. S. & Stormo, G. (1981). *Annu. Rev. Microbiol.* **35**, 365–403.

Jensen, D. E., Kelly, R. C. & von Hippel, P. H. (1976). *J. Biol. Chem.* **251**, 7215–7228.

Kao-Huang, Y., Revzin, A., Butler, A. P., O'Connor, P., Noble, D. & von Hippel, P. H. (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 4228–4232.

Karam, J., Gold, L., Singer, B. S. & Dawson, M. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 4669–4673.

Kelly, R. C., Jensen, D. E. & von Hippel, P. H. (1976). *J. Biol. Chem.* **251**, 7240–7250.

Kowalczykowski, S. C., Lonberg, N., Newport, J. W. & von Hippel, P. H. (1981). *J. Mol. Biol.* **145**, 75–104.

Krisch, H. M. & Allet, B. (1982). *Proc. Nat. Acad. Sci., U.S.A.* **79**, 4937–4941.

Krisch, H., Bolle, A. & Epstein, R. H. (1974). *J. Mol. Biol.* **88**, 89–104.

Krisch, H. M., Duvoisin, R. M., Allet, B. & Epstein, R. H. (1980). *ICN-UCLA Symp. Mol. Cell Biol.* **19**, 517–526.

Lemaire, G., Gold, L. & Yarus, M. (1978). *J. Mol. Biol.* **126**, 73–90.

Lonberg, N., Kowalczykowski, S. C., Paul, L. S. & von Hippel, P. H. (1981). *J. Mol. Biol.* **145**, 123–138.

McGhee, J. D. & von Hippel, P. H. (1974). *J. Mol. Biol.* **86**, 469–489.

Newport, J. W., Kowalczykowski, S. C., Lonberg, N., Paul, L. S. & von Hippel, P. H. (1981*a*). *ICN-UCLA Symp. Mol. Cell Biol.* **19**, 485–505.

Newport, J. W., Lonberg, N., Kowalczykowski, S. C. & von Hippel, P. H. (1981*b*). *J. Mol. Biol.* **145**, 105–121.

Noller, H. F. (1980). In *Ribosomes: Structure, Function and Genetics* (Chambliss, G., Craven, G. R., Davies, J., Davis, K., Kahan, L. & Nomura, M., eds), pp. 3–22, University Park Press, Baltimore.

Nomura, M., Yates, J. L., Dean, D. & Post, L. E. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 7084–7088.

Nomura, M., Dean, D. & Yates, J. L. (1982). *Trends Biochem. Sci.* **7**, 92–95.

Nussinov, R. & Jacobson, A. B. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 6309–6313.

Revzin, A. & von Hippel, P. H. (1977). *Biochemistry*, **16**, 4769–4776.

Russel, M. L., Gold, L., Morrissett, H. & O'Farrell, P. Z. (1976). *J. Biol. Chem.* **251**, 7263–7270.

Schneider, T. D., Stormo, G. D., Haemer, J. & Gold, L. (1982). *Nucl. Acids Res.* **10**, 3013–3024.

Spahr, D. F., Farber, M. & Gesteland, R. F. (1969). *Nature (London)*, **222**, 455–458.

Stormo, G. D., Schneider, T. D. & Gold, L. (1982). *Nucl. Acids Res.* **10**, 2971–2996.

Tinoco, I., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M. & Gralla, J. (1973). *Nature New Biol.* **246**, 40–41.

von Hippel, P. H. (1979). In *Biological Regulation and Development* (Goldberger, R. F., ed.), vol. 1, pp. 279–347, Plenum Publishing Corp., New York.

von Hippel, P. H., Revzin, A., Gross, C. A. & Wang, A. C. (1974). *Proc. Nat. Acad. Sci., U.S.A.* **71**, 4808–4812.

Weber, H., Billeter, M. A., Kahane, S., Weissmann, C., Hindley, J. & Porter, A. (1972). *Nature New Biol.* **237**, 166–170.

Zuker, M. & Stiegler, P. (1981). *Nucl. Acids Res.* **9**, 133–148.

*Edited by M. Gellert*