Volume 6 Number 6 1979 Nucleic Acids Research

Equilibrium and kinetic aspects of protein-DNA recognition

M.A.Livshitz, G.V.Gursky, A.S.Zasedatelev and M.V.Volkenstein

Institute of Molecular Biology, USSR Academy of Sciences, Moscow 117984, USSR

Received 12 March 1979

ABSTRACT

The specificity of regulatory protein binding to DNA is due to a complementarity between the sequence of reaction centres on the protein and the base pair sequence in the specific DNA site allowing the formation of a number of specific noncovalent bonds between the interacting entities. In the present communication the thermodynamic and kinetic aspects of these interactions are considered. The extent of binding specificity is shown to increase with an increase of the bond stability constants and with an increase in the number of ligand reaction centres. Kinetic analysis is carried out assuming that association process is very fast and that dissociation of nonspecific complexes is a rate-limiting step in the recognition of a specific binding site on DNA. The calculations show that a ligand can recognize its specific binding site on DNA within a reasonably limited time interval if the number of its reaction centres and the corresponding stability constants are strongly limited.

INTRODUCTION

At the present stage of investigation of the processes of gene activity regulation it is important to find out the mechanisms responsible for the binding specificity of regulatory proteins for their target sites on DNA and to estimate the thermodynamic and kinetic parameters of the binding processes. In the early models for the protein-DNA recognition it was supposed that the conformation of regulatory sites on DNA differs from that of the other DNA sites and that regulatory proteins recognize these differences[1-3]. It was also supposed that the binding of a protein to DNA leads to local unwinding of the DNA helix at the specific binding site[4]. Now it is well established that the control sites on DNA have no special conformational properties and that the three principal types of regulatory proteins- repressors, RNA-polymerases and restrictases are able to recognize specific nucleo-

tide sequences in double helical DNA without disrupture of the DNA structure[5-9]. Therefore, it seems plausible that the recognition is based on the direct correspondence between the sequence of AT- and GC-specific reaction centres on the protein surface and base pair sequence in the corresponding control site on DNA. The most advantageous conditions for protein binding are realized when the sequence of specific protein reaction centres is strictly complementary to the base pair sequence at the control site on DNA (Fig.1).

If the protein binds to the site partially overlapped with the specific binding site the association constant is reduced due to a loss of a certain number of specific contacts. The model presented in Fig.1 has been used to calculate the adsorption isotherms for multisite ligands each covering several consecutive residues on a heteropolymer[10-13] and to consider stereochemical aspects of specific binding of regulatory proteins to DNA[14-16].

In the present paper we shall consider thermodynamic and kinetic aspects of the recognition problem, formulate the criteria allowing to evaluate the extent of ligand binding specificity and estimate the time needed for a ligand to find its specific binding site.

THERMODYNAMICS OF SPECIFIC BINDING

Suppose the ligand is two-component i.e. it contains L_1 AT-

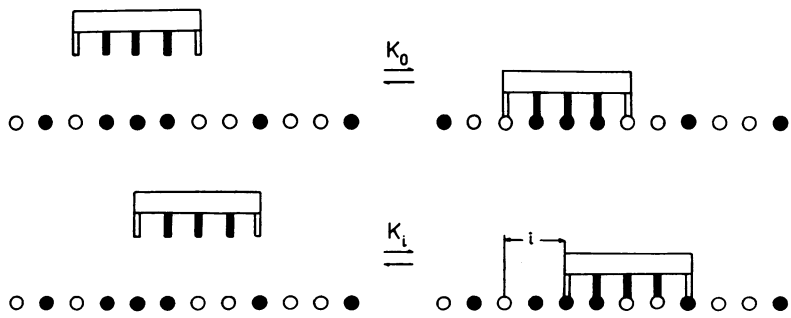
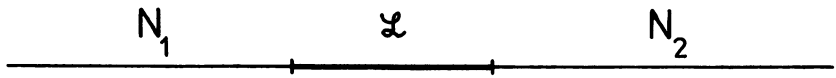


Fig.1. A schematic representation of binding reactions of a protein with its specific binding site on DNA and with a site shifted from the specific binding site by i = 2 base pairs. K_0 and K_i are the corresponding equilibrium association constants. Open and full circles symbolize AT and GC base pairs. Vertical bars represent the protein reaction centres.

-pair specific reaction centres and L_2 GC-specific centres. The exchanges AT $\leftrightarrow$ TA and GC $\leftrightarrow$ CG will be supposed to have no influence on the binding constant. (The results can be easily extended to the case when AT $\neq$ TA, GC $\neq$ CG upon binding). To take into account the effects due to the physical dimensions of the regulatory protein molecule we shall suppose the ligand to cover and make inaccessible for binding $\mathcal{L}$ successive base pairs of DNA. Specificity and stability of the ligand-DNA complex is due to formation of a number of bonds between the ligand reaction centres and the bases. Let the sequence of AT and GC base pairs in the specific binding site be strictly complementary to the sequence of reaction centres of the ligand (although in real systems it is not always so). Let us consider an ensemble of the DNA molecules with a given base pair sequence being in thermodynamic equilibrium with the ligand. Every DNA molecule contains a specific ligand binding site surrounded by long sections with a random base pair sequence:



specific site

Let N_1 and N_2 be the numbers of base pairs in these sections ($N_1 \gg L$, $N_2 \gg L$). For accurate recognition the specific base pair sequence ought to be unique within the polymer. This implies that the specific sequence should be sufficiently long so that the probability of finding the same sequence in the rest of the polymer would be negligible. This is equivalent to the requirement that

$$(N_1 + N_2)\, x_1^{L_1} x_2^{L_2} \ll 1 \qquad (1)$$

where $x_1 (x_2)$ is the probability of finding an AT (GC) pair in the random sequence sections of the polymer ($x_1 + x_2 = 1$). Analogous condition in the case of "four-letter recognition" is:

$$(N_1 + N_2) \prod_{\alpha=1}^{4} x_\alpha^{L_\alpha} \ll 1 \qquad (2)$$

here x_1, x_2, x_3 and x_4 are the probabilities of finding the AT, TA, GC and CG base pairs, correspondingly ($\sum_{\alpha}^{4} x_\alpha = 1$), at a given site on DNA. L_1, L_2, L_3 and L_4 are the numbers of AT-, TA-, GC- and CG-specific reaction centers of the ligand. The inequalities (1) and (2) establish certain relationships between the length of the polymer and the number of ligand reaction centres allowing to consider the specific ligand binding site as being unique on the polymer. These relationships permit one to compare the recognition capacities of two- and four-component ligands in terms of economy of genetic material. Let the frequency of all types of base pairs in the DNA sequences lying outside the specific binding site be the same and let $L_1 = L_2 = L/2$ for a two-component ligand and $L_1 = L_2 = L_3 = L_4 = L/4$ for a four-component ligand. Then (1) and (2) turn into:

$$(N_1 + N_2)(1/2)^L \ll 1 \qquad (3)$$

$$(N_1 + N_2)(1/4)^L \ll 1 \qquad (4)$$

From these relations one can conclude that the number of ligand reaction centres, L, which satisfies the condition (4) is approximately twice as small as compared with the value found from the condition (3). This means that a four-letter recognition code is more economic than a two-letter code.

The thermodynamic properties of binding are fully described by the grand partition function [17]:

$$\Xi = \sum_q \lambda^q Z_q \qquad (5)$$

Here λ is the absolute activity of the ligand, Z_q is a canonical partition function for a polymer (of a given base pair sequence) with q ligand molecules adsorbed. If the sequences of N_1 and N_2 base pairs are random, the mean number of the bound ligand molecules per base pair, r, is:

$$r = \frac{1}{N_1 + N_2 + \mathcal{L}} \frac{\partial \langle \ln \Xi \rangle}{\partial \ln \lambda} \qquad (6)$$

Here the angular brackets stand for the averaging over all possible base pair sequences outside the specific site. The specific site sequence remains fixed.

As far as the concentration of regulatory proteins in the cell is small the activity λ in (6) can be replaced by the free li-

gand concentration m, and only the first terms of the power series expansion of $\bar{z}$ are essential. Such an approximation leads to:

$$\lim_{r\to 0} \frac{r}{m} = \frac{\langle z_1 \rangle}{N_1+N_2+\varkappa} = \frac{N_1+N_2-2\varkappa+1}{N_1+N_2+\varkappa} \langle K \rangle + \sum_{0<|i|<\varkappa} \langle K_i \rangle + K_0 \qquad (7)$$

where $\langle K \rangle$ is the mean equilibrium constant for the binding of a ligand to nonspecific DNA sections. Averaging is carried out over all possible base pair sequences lying outside the specific binding site. $\langle K_i \rangle$ is the mean association constant of the ligand with a site shifted from the specific binding site by i base pairs. If $|i| < \varkappa$ this site is partially overlapped with the specific binding site (see Fig.1). K_0 is the binding constant of the ligand to the specific binding site. The probability of finding the ligand in the bound state at the specific binding site is

$$R = \frac{K_0}{(N_1+N_2-2\varkappa+1)\langle K \rangle + \sum_{0<|i|<\varkappa} \langle K_i \rangle + K_0} \qquad (8)$$

The R value can be regarded as a measure of selectivity of ligand binding. To calculate R one must evaluate K_0, $\langle K \rangle$ and $\langle K_i \rangle$.

EVALUATION OF BINDING CONSTANT K_0.

The constant K_0 can be represented as a sum of statistical weights of various ligand adsorption states. Any adsorption state is specified by the indication of those reaction centres which are bound to DNA base pairs and of those which are not. Let the variable θ_i specifies the state of the i-th ligand reaction centre: $\theta_i = 1$ if the i-th centre is bound, and $\theta_i = 0$ if it is not. The ligand reaction centres can form the bonds either independently of each other or can interact with DNA base pairs in a cooperative manner. In the latter case we shall suppose the state of a reaction centre to depend on the states of its two nearest neighbours. Let the reaction centres form an unhyphenated sequence of length $L = L_1 + L_2$. The constant K_0 is calculated as a partition function for a finite two-state Ising lattice[18]:

$$K_0 = -1 + \sum_{\{\theta\}} \exp\left(-\frac{\Delta F(\theta)}{RT}\right) \qquad (9)$$

Here $\Delta F(\theta)$ is the free energy of the adsorption state speci-

fied by a given set $\{\theta\} = (\theta_1, \theta_2 \ldots \theta_L)$. The summation is carried out over all possible $\theta_i = 0,1$. The unity is subtracted to eliminate the state in which all ligand reaction centres are nonbonded ($\theta_i = 0$, $1 \leq i \leq L$). Since any state of a ligand reaction centre can be correlated with the states of its two nearest neighbours, the free energy of a ligand-polymer complex can be expressed as:

$$\Delta F(\theta) = -\sum_{j=1}^{L} U_j \theta_j + V \sum_{j=2}^{L} (\theta_j - \theta_{j-1})^2 \qquad (10)$$

where $U_j = RT \ln S^{(j)}$ is the free energy change accompanying the formation of a bond between the j-th ligand reaction centre and DNA base pair, provided that the j-th reaction centre lies at the interior of an uninterrupted sequence of bonded reaction centres, $S^{(j)}$ is the corresponding stability constant. $V = -\frac{1}{2} RT \ln \sigma$ is the free energy change associated with the formation of a boundary between adjacent stretches of bonded and nonbonded ligand reaction centres ($V \geq 0$). The free energy change V takes into account so-called "strain energy" arising on the boundaries between bonded and nonbonded ligand reaction centres. The constant σ is the cooperativity parameter ($\sigma = 1$ stands for the absence of cooperativity, i.e. independent binding of reaction centres; $\sigma \to 0$ stands for the high cooperativity of binding, i.e. "all-or-none" binding).

From Eq (9) and Eq (10) it follows that

$$K_o = -1 + \sum_{\{\theta\}} S^{(1)\theta_1} \prod_{j=2}^{L} S^{(j)\theta_j} (\sqrt{\sigma})^{(\theta_j - \theta_{j-1})^2} \qquad (11)$$

which can be represented in a matrix form:

$$K_o = -1 + (1,1)\left(\prod_j \mathbb{M}_j\right)\begin{pmatrix} 1 \\ S^{(1)} \end{pmatrix} \qquad (12)$$

where

$$\mathbb{M}_j = \begin{pmatrix} 1 & \sqrt{\sigma} \\ S^{(j)}\sqrt{\sigma} & S^{(j)} \end{pmatrix} \qquad (13)$$

$S^{(j)} = S_1$ or $S^{(j)} = S_2$ depending on whether the j-th reaction centre is specific to AT or to GC pair. In a special case of only one type of reaction centres ($S^{(j)} = S$) Eq(12) takes the following form:

$$K_0 = -1 + (1,1)\mathbf{M}^{L-1}\begin{pmatrix}1\\ s\end{pmatrix} \qquad (14)$$

From Eqs (13) and (14) by matrix algebra methods one finds that

$$K_0 = \chi_1^L + \chi_2^L - 1 + 2s\sqrt{\sigma}(1-\sqrt{\sigma})\frac{\chi_1^{L-1} - \chi_2^{L-2}}{\chi_1 - \chi_2} \qquad (15)$$

where χ_1 and χ_2 are the eigenvalues of the matrix M:

$$\chi_{1,2} = \frac{s+1}{2} \pm \frac{1}{2}\sqrt{(s-1)^2 + 4s\sigma} \qquad (16)$$

In the case of independent reaction centres ($\sigma \to 1$) this gives:

$$K_0 = (s+1)^L - 1 \qquad (17)$$

In the opposite case of high cooperativity ($\sigma \to 0$) Eq (15) gives:

$$K_0 = s^L \qquad (18)$$

Such all-or-none type of binding can be due not only to a high cooperativity of binding reaction ($\sigma \to 0$) but also can take place if the stability constant $s \gg 1$ (in that case Eq (17) reduces to Eq (18).

EVALUATION OF BINDING CONSTANT $<K>$ AND $<K_1>$.

The binding constant for a random base pair sequence DNA $<K>$ can be calculated by averaging (9) over all possible base sequences. It can be found that

$$<K> = -1 + (1,1)\left(\prod_j <M_j>\right)\begin{pmatrix}1\\ <s^{(j)}>\end{pmatrix} \qquad (19)$$

where

$$<M_j> = \begin{pmatrix}1 & \sqrt{\sigma}\\ <s^{(j)}>\sqrt{\sigma} & <s^{(j)}>\end{pmatrix} \qquad (20)$$

with $<s^{(j)}> = <s_1>$ if the j-th reaction centre is AT-specific and $<s^{(j)}> = <s_2>$ if the j-th reaction centre is GC-specific. The average stability constant $<s_\alpha>$ is given by

$$<s_\alpha> = \sum_\beta x_\beta s_{\alpha\beta} \qquad (21)$$

where $s_{\alpha\beta}$ is the stability constant of a contact between the α-type reaction centre of the ligand and β-type base pair of DNA. x_β is the fraction of β-type base pairs in the DNA parts with a random base pair sequence. A value $s_{\alpha\beta} < 1$ ($\alpha \neq \beta$) is assigned for a repulsion between the ligand reaction centre of type α and DNA base pair of type β. For the binding of one-component ligand one finds that

$$\langle K \rangle = \bar{\chi}_1^L + \bar{\chi}_2^L - 1 + 2\langle s \rangle \sqrt{\sigma}(1-\sqrt{\sigma}) \frac{\bar{\chi}_1^{L-1} - \bar{\chi}_2^{L-1}}{\chi_1 - \chi_2} \qquad (22)$$

where

$$\bar{\chi}_{1,2} = \frac{\langle s \rangle + 1}{2} \pm \frac{1}{2}\sqrt{(\langle s \rangle - 1)^2 + 4\sigma\langle s \rangle} \qquad (23)$$

For a two-component ligand $\langle K \rangle$ depends on the distribution pattern of AT and GC-specific reaction centres along the ligand. However, very simple expressions for $\langle K \rangle$ can be obtained in the limiting cases of noncooperative ($\sigma \to 1$) and highly cooperative ($\sigma \to 0$) binding mechanisms:

$$\langle K \rangle = \begin{cases} \langle s_1 + 1 \rangle^{L_1} \langle s_2 + 1 \rangle^{L_2} - 1 & \text{(independent binding)} \\ \langle s_1 \rangle^{L_1} \langle s_2 \rangle^{L_2} & \text{(all-or-none binding)} \end{cases} \qquad (24)$$

Further consideration will be carried out for these two limiting cases.

To evaluate the average association constant of a ligand with a site partially overlapped with the specific binding site one should take into account that in this case the ligand reaction centres can be divided into two groups depending on whether they interact with base sequences lying beyond the specific binding site or form specific contacts with bases at the specific binding site. The contribution of the latter type of reaction centres to the overall binding free energy can easily be evaluated for any specified sequence of ligand reaction centres. Let the ligand be shifted by i ($|i| < \mathcal{L}$) base pairs from the specific binding site and let $q_{\alpha\beta}(i)$ denote the number of coincidences of the ligand reaction centres of type α ($\alpha = 1,2$) with base pairs of type β ($\beta = 1,2$) at the specific binding site. Here $\alpha = 1,2$ stands for the AT and GC=specific reaction centres, respectively; $\beta = 1,2$ stands for AT and GC base pairs. The numbers $q_{\alpha\beta}(i)$ can readily be calculated for any given arrangement of reaction centres along the ligand and for any $|i| < \mathcal{L}$. For example, from Fig.1 we can find that for $i = 2$ $q_{11}(i)=0$, $q_{12}(i)= 2$, $q_{22}(i) = 1$ and $q_{21}(i)= 2$. With $q_{\alpha\beta}(i)$ being calculated $\langle K_i \rangle$ is given by

$$\langle K_i \rangle = (s_1 + 1)^{q_{11}(i)} (\hat{s}_1 + 1)^{q_{12}(i)} (s_2 + 1)^{q_{22}(i)} (\hat{s}_2 + 1)^{q_{21}(i)} \times \langle s_1 + 1 \rangle^{L_1 - q_{11}(i) - q_{12}(i)} \langle s_2 + 1 \rangle^{L_2 - q_{21}(i) - q_{22}(i)} \qquad (25)$$

for the noncooperative binding case ($\sigma = 1$) and

$$\langle K_i \rangle = s_1^{q_{11}(i)} \hat{s}_1^{q_{12}(i)} s_2^{q_{22}(i)} \hat{s}_2^{q_{21}(i)} \langle s_1 \rangle^{L_1 - q_{11}(i) - q_{12}(i)} \langle s_2 \rangle^{L_2 - q_{21}(i) - q_{22}(i)} \quad (26)$$

for the binding in an all-or-none manner.

THE ESTIMATION OF THE EXTENT OF LIGAND BINDING SPECIFICITY.

As is seen from Eqs(8) and (24)-(26) the extent of binding specificity exhibited by a ligand at low limit of binding increases with the increase of s_1 and s_2. Let us consider the binding of an one-component ligand carrying an uninterrupted sequence of AT-specific reaction centres ($L_1 = L = \mathcal{L}$, $L_2 = 0$). In this case $q_{11}(i) = L - |i|$ for $|i| < L$; $q_{11}(i) = 0$ for $|i| \geq L$; $q_{12}(i) = q_{21}(i) = q_{22}(i) = 0$ and

$$R = \frac{(s+1)^L - 1}{(s+1)^L - 1 + (N_1 + N_2 - 2L + 2)(\langle s+1 \rangle^L - 1) + 2\,\dfrac{(s+1)^L \langle s+1 \rangle - \langle s+1 \rangle^L (s+1)}{(s+1) - \langle s+1 \rangle} - 2(L-1)} \quad (27)$$

If the binding proceeds in an all-or-none manner (either $\sigma \to 0$ or $s \gg 1$), R is given by the relation (28):

$$R = \frac{s^L}{s^L + (N_1 + N_2 - 2L + 2)\langle s \rangle^L + 2\,\dfrac{s^L \langle s \rangle - \langle s \rangle^L s}{s - \langle s \rangle}} \quad (28)$$

which for $s_{1,2} \to 0$ (strong repulsion) reduces to

$$R_{max} = \frac{1}{1 + (N_1 + N_2 - 2L + 2)x^L + 2\,\dfrac{x - x^L}{1 - x}} \quad (29)$$

The extent of binding specificity strongly depends on the number of ligand reaction centres. For example, if $N_1 + N_2 = 10^6$, $x = 0.5$ and $\mathcal{L} = L$, then $R_{max} \simeq 10^{-3}$ for $L = 10$ and $R_{max} \simeq$ $\simeq 0.25$ for $L = 20$. As L tends to infinity $R_{max} \to R_\infty = (1-x)/(1+x)$ This implies that the extent of specificity for the binding of one-component ligand is strongly limited and cannot exceed $(1-x)/(1+x)$. In Fig.2 the dependences of R on S are shown for the binding of one- and two-component ligands to a two-component DNA. Calculations are carried out for $N_1 + N_2 = 10^6$, $x_1 = x_2 =$ $= 0.5$ and for L ranging from 10 to 30. For the purpose of a comparison in Fig.2 the curves are also shown for the binding of four-component ligand to a four-component DNA. In these calcula-

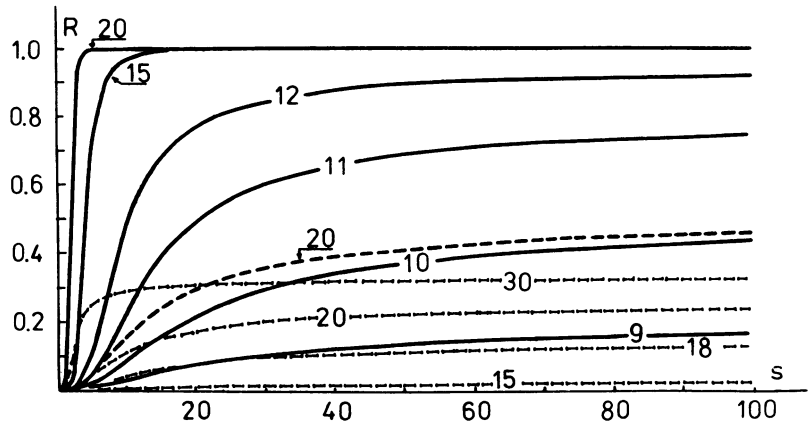


Fig.2. The extent of binding specificity, R , versus the stability constant S for various values of L . (– – –) - one-component ligand; (——) - two-component ligand; (————) - four-component ligand. The total number of ligand reaction centres is indicated by each curve.

tions we assumed that $x_\beta = 1/4$, $L_\beta = L/4$ and $S_{\alpha\beta} = \begin{cases} S & \alpha = \beta \\ 0 & \alpha \neq \beta \end{cases}$ where $\alpha, \beta = 1,2,3,4$. Similar assumptions were made in the case of two-component ligand binding.

It should be noted that in the case of binding of a two-(four)--component ligand the probability of binding to DNA sites partially overlapped with the specific binding site depends on the arrangement of reaction centres along the ligand. This dependence, however, can be neglected in the most practical cases. Clearly, the probability of ligand binding to DNA sites overlapping with the specific binding site takes the greatest value in the case of binding of one-component ligand with an uninterrupted sequence of reaction centres. Indeed, shifting such a ligand from the specific binding site by one base pair may lead to a loss of only one specific contact, while in the case of binding of a two-component ligand such a shift results in the loss of several specific contacts. As a consequence, the probability of binding of two- or four-component ligand to DNA sites partially overlapped with the specific binding site is much lower than the probability of nonspecific binding:

$$\sum_{0<|i|<\infty} \langle K_i \rangle \ll (N_1 + N_2) \langle K \rangle \qquad (31)$$

(In the case of binding of an one-component ligand these probabilities can be comparable).

Fig.3 shows that the accuracy of recognition of a specific site within a heterogeneous DNA strongly depends on the polymer length.

From Eqs.(8) and (24)-(26) it follows that the greatest extent of binding specificity is achieved if $q_{11}(i)$ and $q_{22}(i)$ take their lowest values. This means that for optimal recognition the sequence of ligand reaction centres must belong to a class of uncorrelated sequences. For such sequences the condition (31) is always fullfilled.

The extent of binding specificity exhibited by two- or four--component ligands can be calculated with a good accuracy from the relation:

$$R \simeq \frac{K_0}{(N_1+N_2)\langle K \rangle + K_0} \qquad (32)$$

In the case of binding of a two-component ligand $\langle K \rangle$ is given by Eq.(24). For the four-component ligand case the analog of Eq.

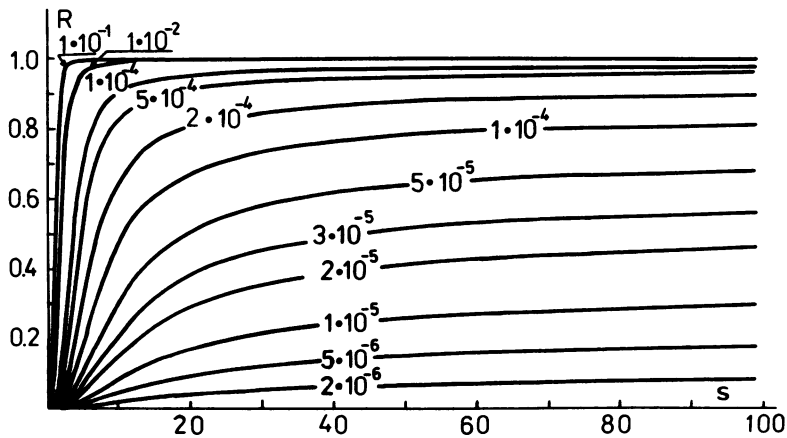


Fig.3. The binding specificity R of a two-component ligand versus the stability constant s for various values of $L/(N_1+N_2)$. The total number of ligand reaction centres $L = 20$.

(24) is

$$\langle K\rangle = \begin{cases} -1 + \prod_{\alpha=1}^{4} \langle 1 + S_\alpha \rangle^{L_\alpha} & \sigma = 1 \\ \prod_{\alpha=1}^{4} \langle S_\alpha \rangle^{L_\alpha} & \sigma = 0 \end{cases} \qquad (33)$$

where $\langle S_\alpha \rangle = \sum_{\beta=1}^{4} S_{\alpha\beta} x_\beta$

As is seen in Fig.4, the R value is higher the lower are the stability constants for unfavourable contacts $S_{\alpha\beta}$ $(\alpha \neq \beta)$. Earlier[12] we have suggested a criterium allowing one to estimate the extent of binding specificity exhibited by a ligand

$$\frac{K_0}{(N_1+N_2)\langle K\rangle} \geq 1 \qquad (34)$$

If the probability of ligand binding to DNA sites overlapped with the specific binding site is negligible in comparison with the probability of nonspecific binding (See Ineq.(31)) the condition (34) is equivalent to the requirement that $R > 1/2$. The quantitative estimations of the parameters for which $R \simeq 1/2$ can be obtained from Figs.2 and 3.

For $N_1 + N_2 = 10^6$, $x_\alpha = 1/4$ $S_{\alpha\alpha} = 30$ and $S_{\alpha\beta} = 0 (\alpha \neq \beta)$ a four-component ligand provides $R = 1/2$ if $L = 10$. A two-component ligand must have 20 reaction centres to manifest the same extent of binding specificity under these conditions. In

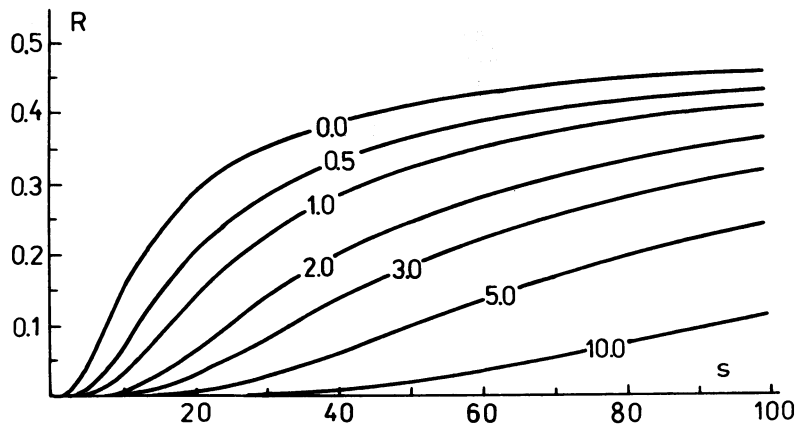


Fig.4. The dependences of R on S calculated for the binding of a two-component ligand. Curves are shown for various values of stability constant $S_{\alpha\beta} = \hat{S}$ $(\alpha \neq \beta)$. Calculations are carried out for $N_1 + N_2 = 10^6$ and $L = 20$.

the case of binding of an one-component ligand $R = 1/2$ cannot be achieved even with $S \to \infty$ and $L \to \infty$. In real regulatory proteins the lattice of reaction centres appears to be a two-component one for stereochemical reasons but AT $\neq$ TA and GC $\neq$ CG upon binding[16].

KINETICS OF SPECIFIC COMPLEX FORMATION.

In previous section it is demonstrated that specificity of binding increases with the increase of L and s values. However, the increase in specificity is accompanied by an increase in the strength of ligand binding to nonspecific binding sites as well. Nonspecific binding competes with specific binding and plays an important role in the kinetics of ligand binding to a specific binding site. In the context of recognition problem the dissociation rate of nonspecific complexes is of particular importance. As is well known the regulatory protein - DNA association process is very fast[19,20]. Dissociation from nonspecific binding sites is a much slower process[21]. Clearly, the time needed for a protein to find its specific binding site on DNA should be reasonably limited. If a regulatory protein searches out its specific binding site on DNA by random collisions coupled with a series of association and dissociation processes then strong constraints must be imposed on the dissociation rate of nonspecific complexes.

Let there be many nonspecific adsorption sites $n = \sum n_j$ differing in the magnitude of dissociation rates $1/\tau_j$ for the ligand. And let there also be $n^* \ll n$ specific sites (such as operators) with a much longer release time $\tau^* \gg \tau_j$. The ligand molecules will then tend to be accumulated in these most favourable adsorption sites. The kinetics of association and dissociation processes at a low level of binding is described by the following equations:

$$\begin{aligned} \frac{dP_j}{dt} &= k n_j P_0 - P_j/\tau_j \\ \frac{dP_0}{dt} &= \sum_j P_j/\tau_j - k n P_0 \\ \frac{dP^*}{dt} &= k n^* P_0 - P^*/\tau^* \end{aligned} \qquad (35)$$

with $P_0(0)=1$, $P_j(0) = 0$, $P^*(0) = 0$

Here P_0 is the probability of finding a ligand in a free state, P^* and P_j refer to the probabilities of finding the ligand complexed with the specific operator site and the nonspecific j-th site, respectively $P_0 + \sum_j P_j + P^* = 1$.

We suggest that the rate-limiting step of association process (in specific as well as in nonspecific binding) is diffusion controlled with the characteristic rate constants (k) of about $10^8 \div 10^{10}$ M^{-1} sec^{-1}. The first two equations in (35) describe relatively fast processes, while the last equation corresponds to the slowest process kn^*, $1/\tau^* \ll kn_j$, $1/\tau_j$; Therefore after a short transition period the fast variables attain their stationary values

$$P_j \to kn_j\tau_j P_0 , \quad P_0 \to (1-P^*)/(1+k\sum_j n_j\tau_j)$$

and then change slowly with P^*

$$\frac{dP^*}{dt} = \frac{kn^*}{1+kn\langle\tau\rangle}(1-P^*) - P^*/\tau^* \qquad (36)$$

Here the mean releasing time $\langle\tau\rangle = \frac{1}{n}\sum_j n_j\tau_j$. The solution of the Eq.(36) is:

$$P^*(t) = \frac{kn^*\tau^*}{kn^*\tau^* + kn\langle\tau\rangle + 1}\left(1 - e^{-t/t_{search}}\right)$$

$$1/t_{search} = \frac{kn^*}{1+kn\langle\tau\rangle} + \frac{1}{\tau^*} \qquad (37)$$

If the lifetime of a specific complex is much longer than all other relevant times ($\tau^* \to \infty$) then equation (37) reduces to

$$P^* = 1 - e^{-t/t_{search}} \qquad (38)$$

where

$$t_{search} = \frac{n}{n^*}\left(\frac{1}{kn} + \langle\tau\rangle\right) \qquad (39)$$

This result has a clear physical meaning. Before the finding a specific binding site on DNA a ligand must visit nearly all n/n^* nonspecific binding sites. Each trial on average proceeds during the time length which can be represented as a sum of two characteristic times: the time $1/kn$ needed for a ligand to encounter a binding site on DNA and the average time for the release of the ligand from a nonspecific binding site.

The probability of dissociation of a ligand-polymer complex is dependent on the elementary rate constant, ν , associated with the disruption of a single ligand-polymer bond and on the thermodynamic probability of a state in which all bonds except one are disrupted. So the mean dissociation time for non-specific ligand-polymer complexes is

$$\langle \tau \rangle = \left\langle \frac{1}{\nu} \frac{K}{K(1)} \right\rangle \approx \frac{1}{\nu} \frac{\langle K \rangle}{L s \sigma} \qquad (40)$$

Here K is the statistical weight for all states of the ligand-polymer complex with a given binding site, $K(1)$ is the statistical weight for those states in which only one specific ligand-polymer bond remains intact. $\sigma < 1$ is the cooperativity parameter associated with the ligand reaction centres. The relation (40) is obtained with assumption that the rates of formation and disruption of particular ligand-polymer bonds are sufficiently great with respect to the characteristic dissociation rate of the complex so that various adsorption states of the ligand are equilibrated.

In order to determine the intrinsic second-order rate constant k for the binding of a protein to DNA one needs to use very dilute solutions in which the association process between the protein and DNA is a rate-limiting step ($1/kn \gg \langle \tau \rangle$). The rate of association of lac repressor to lac operator was measured by Riggs et al[19] in very dilute solutions containing about 10^{-12} M of each reactant. If $1/kn \gg \langle \tau \rangle$ under the experimental conditions then Eq.(39) predicts that kinetics of specific site selection should be relatively independent of DNA length. Very recently Goeddel et al[20] have found that rates of association of lac repressor to synthetic operators of about 20 base pairs long ($k = 2 \cdot 10^9$ M^{-1} sec^{-1} at 0.05 M KCl and $k = 1 \cdot 10^9$ M^{-1} sec^{-1} at 0.20 M KCl)are very close to the corresponding values determined by Riggs et al[19] from the binding of the repressor to $\lambda \phi$ 80dlac DNA carrying the lac operator and about 50000 base pairs of non operator DNA ($k = 7 \cdot 10^9$ $M^{-1}sec^{-1}$ at 0.05M KCl and $k = 3 \cdot 10^8$ $M^{-1}sec^{-1}$ at 0.20 M KCl). Since kn is about 350 sec^{-1} under the conditions of this experiment (0.05 M KCl), this may indicate that $\langle \tau \rangle \lesssim 3 \cdot 10^{-3}$ sec, although other explanations are possible. On the other hand, the

rate of dissociation of unspecific complexes can be estimated from the experimentally determined[19,21,22] values for equilibrium constants $K_o = 10^{13}$ M^{-1} and $\langle K \rangle = 10^6$ M^{-1} and the rate of dissociation of repressor-operator complex $1/\tau^* = 6 \cdot 10^{-4}$ sec^{-1}. Assuming that the rates of association are of the same order of magnitude for the binding of the repressor to the operator and nonoperator DNA, one can find that $\langle \tau \rangle \sim 10^{-4}$ sec. Approximately the same estimate for $\langle \tau \rangle$ can be obtained from the kinetics observations of Jobe et al[23] showing that on adding of chicken blood DNA with a concentration of $1.1 \cdot 10^{-5}$ M base pairs there is a decrease in the rate of association of lac repressor to the operator from $7 \cdot 10^9$ $M^{-1}sec^{-1}$ to an apparent rate of $1.2 \cdot 10^9$ $M^{-1}sec^{-1}$. This indicates that the weak binding to non-operator DNA interferes with the search of lac repressor for the operator. Applying Eq (39) one can find from these data $\langle \tau \rangle \sim 10^{-4}$ sec at an ionic strength of 0.05 M. The rate competition observations of Lin andRiggs[22] are consistent with this estimate and further demonstrate that $\langle \tau \rangle$ depends on source of nonoperator DNA and ionic strength of solution. The experimentally determined value for the rate of association of lac repressor to lac operator at low ionic strengths is about 70 times greater than that estimated from the von Smeluchowski's theory for a diffusion-limited reaction. However, the reaction appears to be diffusion controlled since in 20% sucrose its rate is diminished by a factor of two as would be expected from the change in viscosity.[19] Several mechanisms were suggested which could accelerate the search of repressor for its operator. In the first of these models[19,24] the long range electrostatic interactions between the repressor and operator DNA was considered as a rate-enhancing mechanism. However, Richter and Eigen[24] have estimated that electrostatic attraction alone is not sufficient to explain the high rapidity of the binding reaction. In the second model[19,24] it is assumed that unspecific binding of repressor to DNA is accompanied by linear diffusion ("sliding") of the repressor along the DNA chain. This would increase the effective range of the specific binding site thereby increasing the chance for the represser to find the operator. Another mechanism for the searching process through

direct transfer of the repressor from site to site on the DNA without dissociation to free ligand was suggested by von Hippel et al[25]. Although the origin for the high rapidity of the binding reaction is not yet established, the prediction of the last model that the association rate should strongly decrease for small DNA fragments is not supported by kinetics measurements of Goeddel et al.[20] The "sliding" model is considered in detail by a number of investigators.[26-29] In our present paper the possibility for a linear diffusion is not taken into account since we are mainly interested in the limiting case when dissociation from unspecific binding sites is a rate - determining step in the formation of specific protein - DNA complexes. From the point of view of our multipoint attachment model there is no reason to believe that a time needed for one step in the random walk along the DNA is much shorter than the dissociation to free ligand. Clearly the disrupture of all specific bonds between ligand and DNA is required to make possible sliding as well as three-dimensional diffusion of a ligand.

If the total number of nonspecific binding sites is sufficiently large one can neglect $1/Kn$ as compared with $\langle\tau\rangle$ and set in Eq. (39) $t_{search} \simeq \frac{n}{n^*}\langle\tau\rangle$ Under these conditions the rate of specific site selection depends on the DNA length and is proportional to the ratio of unspecific binding sites to specific sites. This type of kinetics was observed by Hinkle and Chamberlin[30] for the promoter sites recognition by RNA polymerase on T7 phage DNA. These authors have found that at DNA concentrations as low as $2{,}6\cdot10^{-10}$ M and at enzyme/DNA ratios ranged from 0.05 to 0.25 the binding reaction between RNA polymerase and tight binding sites on T7 DNA exhibits pseudo first--order kinetics with an apparent rate constant of $3\cdot10^{-2}$ sec^{-1}. They have concluded that the rate-determining step in the formation of specific RNA polymerase-DNA complexes is the release of the enzyme from nonspecific binding sites on DNA.

This mechanism reflects probably the real situation in vivo since in E.coli cells there are several molecules of lac repressor and RNA polymerase (~ 10) per chromosome. This indicates that concentrations of reactants in bacterial cell ($\sim 10^{-2}$ M base pairs for DNA and $\sim 10^{-8}$ M for lac repressor) are much

higher than those used in the kinetics experiments of Riggs et al.[19] showing that one can neglect $1/Kn$ as compared with $\langle\tau\rangle$.

From Eqs (39) and (40) one can estimate which constraints should be imposed on parameters L and s for effective recognition by a protein of specific binding sites on DNA. It seems clear that t_{search} in Eq.(38) must be reasonably limited, e.g. $t_{search} < 100$ sec. It has already been mentioned that for accurate recognition of a specific binding site on DNA molecule containing 10^6 base pairs a two-component ligand should have the number of reaction centres L, equal to 20 or more. If one takes in Eq.(39-40) $L = 20$, $N_1 + N_2 = 10^6$, $\sigma = 1$, the stability constant s must be less than 4 even when $\nu = 10^{13}$ sec^{-1}, which is the highest possible value for the elementary rate constant. The analysis of binding equilibria for this case ($s = 4$, $L = 20$) shows that such weak bonds provide rather low extent of binding specificity. Indeed, from (25) and (27) it follows that $R \sim 3 \cdot 10^{-2}$ thereby indicating that only about three per cent of the ligand molecules is complexed at the specific binding sites under these conditions. On the other hand the comparative study of influence of the single base pair substitutions in the lac-operator on the lac-repressor binding has shown that any substitution reduces the binding constant under physiological conditions by about 30 times.[31] This may indicate that the magnitude of stability constant s agrees with the energy of a hydrogen bond formation ($U \approx 2$ kcal/mole) provided that the repressor reaction centres act independently of each other. An analogous estimate of the binding energy per a reaction centre is obtained for the binding of AT-specific antibiotic distamycin which may serve as the simplest model for the repressor[32]. Electrostatic interactions play an important role in the binding of regulatory proteins to DNA. A general way for evaluation of the electrostatic component of the binding from the experimental dependences of binding constant on the ionic strength has been developed recently by Record et al[33]. Quantitative dependences of the association rate on the ionic strength in "sliding" model were derived by Berg & Blomberg.[26] In the framework of our model the electrostatic interactions can be incorporated by introducing a set of stability constants

which characterize the interactions between positively charged residues on the protein surface and negatively charged DNA phosphate groups. It can be shown that electrostatic interaction included equally in specific as well as in nonspecific binding constants will not have marked influence on our estimates.

The kinetic constraints are so strong that the existence of long multisite ligands ($L \sim 20$) exhibiting a high extent of binding specificity($R \sim 1/2$) seems to be unrealistic. The opposite requirements imposed by formula (34), (38), (39) and (40) can be reconciled if a long multisite ligand is divided into several small parts which upon binding may act to some extent independently. The loosely coupled oligomeric structure of regulatory proteins meets these requirements and may serve as a factor favouring the solution of these kinetic and the equilibrium problems.

REFERENCES

1. Yarus,M.(1969) Ann.Rev.Biochem.38,841-880.
2. von Hippel,P.H. and McGhee, J.D. (1972) Ann.Rev.Biochem.41, 231-300.
3. Gierer,A.(1966) Nature 212,1480-1481.
4. Sobell,H.M. (1973) Adv.Genetics 17, 411-490.
5. Wang,J.C.,Barkley,M.D. and Bourgeois,S. (1974) Nature 851, 247-249.
6. Maniatis,T. and Ptashne,M.,(1973) Proc.Natl.Acad.Sci.USA, 70,1531-1535.
7. Müller-Hill,B.,(1975) Prog.Biophys.Mol.Biol.30,227-252.
8. Chamberlin,M.,(1974) Ann.Rev.Biochem. 43, 721-775.
9. Greene, P.J., Poonian,M.S.,Nussbaum,A.L.,Tobias,L.,Garfin, D.E.,Boyer,H.W. and Goodman,H.M. (1975) J.Mol.Biol.99,237--261.
10. Zasedatelev,A.S.,Gursky,G.V. and Volkenstein,M.V.(1973) Stud.Biophys.40,79-83.
11. Gursky,G.V.,Zasedatelev,A.S.,Volkenstein,M.V.(1972),Mol. Biol. 6,479-490.
12. Livshitz, M.A.,Gursky,G.V.,Zasedatelev,A.S.,and Volkenstein M.V.(1976),Stud.Biophys.60,97-104.
13. Gursky,G.V.,and Zasedatelev A.S.,(1978) Biophysics 23,932--946.
14. Gursky,G.V.,Tumanyan,V.G.,Zasedatelev,A.S.,Zhuze,A.L., Grokhovsky,S.L. and Gottikh,B.P. (1975),Mol.Biol.9,635-651.
15. Gursky,G.V.,Tumanyan,V.G.,Zasedatelev,A.S.,Zhuze,A.L.,Grokhovsky,S.L. and Gottikh,B.P.(1976),Mol.Biol.Rep.2,413-425.
16. Gursky,G.V.,Tumanyan,V.G.,Zasedatelev,A.S.,Zhuze,A.L.,Grokhovsky,S.L. and Gottikh,B.P.,(1977) in: Nucleic Acid-Protein Recognition,Ed.H.Vogel,189-217.

17. Hill,T.L.(1960) An Introduction to Statistical Thermodynamics,Addison-Wesley,Reading,Mass.
18. Poland,D. & Scheraga,H.A.(1970) Theory of Helix-Coil Transitions in Biopolymers, Academic press,New York.
19. Riggs,A.D., Bourgeois,S., and Cohn,M. (1970), J.Mol.Biol. 53,401-417.
20. Goeddel,D.V., Yansura,D.G.,Carnthers,M.H.(1977) Proc.Natl. Acad.Sci.USA 74,3292-3296.
21. Lin,S.Y. and Riggs,A.D.(1975),Cell 4, 107-111.
22. Lin,S.Y. and Riggs,A.D. (1972),J.Mol.Biol.72,671-690.
23. Jobe,A.,Riggs A.D.,Bourgeois,S.(1972) J.Mol.Biol. 64,181--199.
24. Richter,P.H. and Eigen, M. (1974) Biophys.Chem.2,255-263.
25. von Hippel,P.H., Revzin, A., Gross,C.A. and Wang,A.C.(1975), Protein-Ligand Interactions,Eds.H.Sund and G.Blauer,270-288, Walter de Gruyter & Co, Berlin, New York.
26. Berg O.G., Blomberg C. (1976) Biophys.Chem.4,367-381.
27. Berg O.G.,Blomberg C.(1977) Biophys.Chem.7,33.
28. Schranner Richter P.H.(1978) Biophys.Chem.8,135.
29. Berg O.G.,Blomberg C.(1978) Biophys.Chem.8,271.
30. Hinkle D.C.,Chamberlin M.J.(1972) J.Mol.Biol. 70,187-195.
31. Gilbert,W.,Gralla,J.,Majors,J. and Maxam,A.(1975) in: Protein-Ligand Interactions,Eds.H.Sund and G.Blauer,193-210, Walter de Gruyter & Co, Berlin, New York.
32. Krylov, A.S.,Grokhovsky,S.L.,Zasedatelev,A.S.,Zhuze,A.L., Gursky,G.V.,Gottikh B.P. (1978).Dokl.Acad.Sci.USSR,239,732--739.