Volume 16 Number 11 1988 Nucleic Acids Research

Selection of DNA binding sites by regulatory proteins: the LexA protein and the arginine repressor use different strategies for functional specificity

Otto G.Berg

Department of Molecular Biology, Uppsala University Biomedical Center, PO Box 590, S-75124 Uppsala, Sweden

Received February 29, 1988; Revised and Accepted April 26, 1988

Abstract

The DNA sequences in the operator sites of the arginine regulon and of the SOS regulon have been subject to a statistical analysis. A quantitative correlation is found between the statistics of sequence choice and the activity at individual operator sites in both systems, as expected from theoretical considerations [Berg & von Hippel, J.Mol.Biol. (1987) 193, 723-750]. Based on these correlations it is possible to predict the effect of various sequence mutations. There is a significant difference in the slopes of the correlation lines between sequence and activity for the two systems. From this difference it can be expected that individual point mutations in the ARG boxes will have a much smaller effect on activity than similar changes in the SOS boxes. This difference may be related to a strong cooperative activity at tandem ARG boxes while the binding at SOS boxes appears to be mostly noncooperative.

INTRODUCTION

Gene-regulatory protens bind preferentially to a small set of specific recognition sites among a vast excess of structurally similar competitive sites in the genome. In many cases -- e.g. in the binding of repressor proteins to operator sites -- the biological function is determined primarily by the occupancy (saturation) of the DNA site by the protein. An appropriate saturation level can be achieved by a high specificity with sufficiently strong binding only to the specific recognition sites. Alternatively, it can be achieved by mass action simply providing sufficient protein to bind also a large number of nonspecific sites.

Some repressors, like the lac repressor, control only one operon via one operator site (possibly a secondary operator site can contribute also); others, like the trp repressor, control only a few. The LexA protein and the arginine repressor, in contrast, control several different operons spread throughout the genome providing more than a dozen recognition sites in each case. In these systems there exist sufficient sequence data to allow a a statistical analysis of the specific base-pair choices. Previously we have presented a theory that provides a relation between the statistics of base-pair choice and the level of specific function of the recognition sites. The theory has been developed and applied to the promotors of E. coli (1) and to the binding sites for the cyclic-AMP receptor protein (2).

In this communication the theory is applied to the recognition sites for the LexA protein -- the SOS boxes -- and to the recognition sites for the arginine repressor -- the ARG boxes. It will be shown how the correlations between sequence choice and statistics can provide information on the functional specificity of the recognition sites and also predict the effects of point mutations.

THEORY

Sequence Statistics and Function

The theory (1,2) to be applied below can be summarized as follows. Recognition sequences have been selected in evolution to provide some appropriate binding for the relevant proteins. As a consequence the frequency of base-pair occurrences in the set of recognition sites will carry information about the contribution to binding from individual base pairs. Assuming that individual base pairs in a binding site contribute independently -- at least approximately -- to the binding free energy of a site, it can be shown that the frequency of occurrence $n_{\ell B}$ of a particular base pair B (B= A•T, T•A, G•C, or C•G) at a particular position ℓ in the sites is related to its binding free energy contribution as

$$n_{\ell 0}/n_{\ell B} = \exp(\lambda \varepsilon_{\ell B}) \qquad (1)$$

In this relation $n_{\ell 0}$ denotes the frequency of occurence of the most common (consensus) base pair at position ℓ and $\varepsilon_{\ell B}$ is the reduction in binding free energy (expressed in units of kT and counted >0 for a weaker binding) when a consensus base pair is replaced by B. λ is a selection parameter which is the same for all sites in a set. For a biased genome where A•T and G•C base pairs are not equally common, Eq. (1) must be amended to reflect this bias.

Using Eq. (1) for all base-pairs that are not consensus in a site one can calculate the reduction in binding relative to a site consisting only of consensus base pairs. By summing all the non-consensus contributions, one finds that this reduction, E, for a sequence $\{B_\ell\}$ that uses base pair B_ℓ at position ℓ (ℓ=1,..,s where s is the sequence length or site size) can be expressed as

$$\lambda E = \sum_{\ell=1}^{s} \ell n[(n_{\ell 0}+0.5)/(n_{\ell B_\ell}+0.5)] \qquad (2)$$

The extra terms 0.5 have been introduced as a small-sample correction. While E is a reduction in binding free energy (expressed in units of kT), the combination λE is a purely statistical measure expressing the departure from sequence homology with the consensus sequence; it can be referred to as the heterology index. If K_0 is the binding constant to the best binding (consensus) sequence, the binding constant for a non-consensus sequence can be expressed as $K=K_0 \exp(-E)$. Thus, using Eq. (2) we expect a linear correlation between the logarithm of the binding constant K and the heterology index for a site:

$$\ell nK = \ell nK_0 - (1/\lambda) \sum_{\ell=1}^{s} \ell n[(n_{\ell 0}+0.5)/(n_{\ell B_\ell}+0.5)] \qquad (3)$$

The slope of the correlation line gives the value for the selection parameter λ.

As a measure of the importance of individual positions in the sites, one can calculate the sequence information (3) using the small-sample correction of Berg & von Hippel (2):

$$I_\ell = \sum_B [(n_{\ell B}+1)/(N+4)]\ell n[4(n_{\ell B}+1.5)/(N+4.5)] \qquad (4)$$

Here N is the number of sites in the sample and the sum is taken over the four possible base pairs. When this expression is summed over all positions (ℓ=1,..,s) in the sites, the resulting total sequence information, I_{SEQ}, can be used to estimate the number of pseudosites; the larger the sequence information the smaller is the probability that a binding site will occur at random in the genome [cf. Eq. (8) below].

Repressor Binding and Activity

The binding of a repressor at an operator reduces transcription initiation in proportion to the probability that the site is occupied. If there is no other regulation, the reduction in the gene expression under repressed conditions can be taken as a measure of the repressor-operator binding strength.

Let us assume that there are two operator sites next to each other -- as is the case for most ARG boxes and some of the SOS boxes -- and that binding to either will block transcription initiation. Furthermore, the binding at one site can be assumed to influence the binding strength to the other either cooperatively or anticooperatively. The probability that neither site is occupied is

$$P_0 = 1/(1 + K_1R_f + K_2R_f + wK_1K_2R_f^2) \tag{5}$$

Here, R_f is the concentration of free repressor protein, K_1 and K_2 are the binding constants to the respective binding sites, and w is the cooperativity parameter expressing the increase (if w>1) or decrease (if w<1) in the binding to one site when the other is already bound. If under induced conditions the protein is totally inactive (i.e. R_f=0) the induction ratio (repressibility) will be given by 1/Po.

In the case of a single operator site, the induction ratio minus one is proportional to the binding constant:

$$1/P_0 - 1 = K_1R_f = K_0R_f\exp(-E_1) \tag{6}$$

where the result from Eq. (3) has been introduced in the last equality. In the case of two sites and if cooperative binding dominates (w>>1), the induction ratio will be proportional to the product of the two binding constants:

$$1/P_0 -1 \approx wK_1K_2R_f^2 = w(K_0R_f)^2\exp(-E_1-E_2) \tag{7}$$

In the genome one can expect that there occurs at random a large number of sites that resemble the specific sites in sequence. Since many more sequence combinations can accomplish a weak binding, the number of such randomly occurring pseudosites will increase very rapidly as sites with weaker binding are considered (1,4,5). Thus their influence on protein binding will depend not only on the absolute number of pseudosites resembling the specific ones, but also on how rapidly their numbers increase with decreasing binding strength. The number of protein molecules that are bound at such randomly occurring pseudosites can be calculated as (1):

$$m_s = 2N_TF_AK_sR_f\exp[-I_{SEQ}+(\lambda-1)^2\sigma^2/2] \tag{8}$$

$2N_TF_A$ denotes the number of available sites in the genome -- i.e. sites that are not covered by other proteins or inaccessible for other reasons -- which we have estimated as $6\cdot10^5$ (2). In this expression, $K_S=K_0\exp(-E_{SEQ})$ refers to the binding of an average specific site with heterology index λE_{SEQ}. The sequence information I_{SEQ} is calculated as the sum of the contributions given in Eq. (4). The variance σ^2 can also be calculated from the sequence data (1). Eq. (8) is expected to hold reasonably well if $\lambda\approx1$ or larger.

SOS BOXES

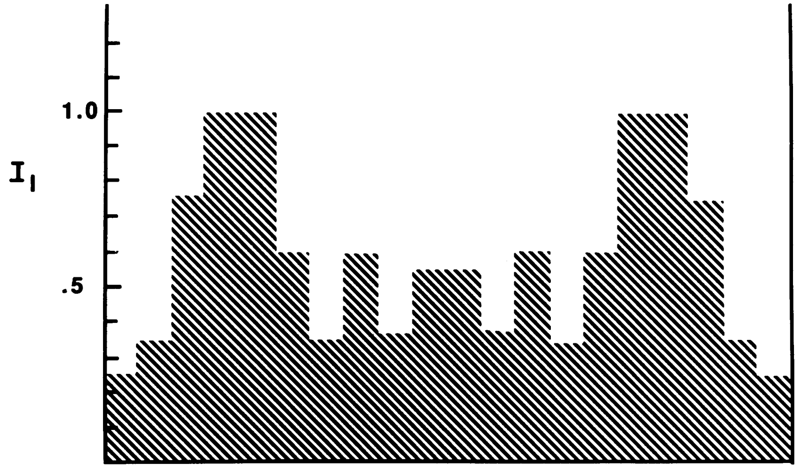
The Basis Set

The LexA protein functions as a repressor for a number of genes involved in the SOS-response (for reviews see 6,7). Around 20 recognition sites for the LexA protein -- the SOS boxes -- have been identified and sequenced. Of the 14 sites listed by Walker (7) we exclude himA which may be nonfunctional (8) and umuDC-1 which does not bind LexA (9). To the remaining ones can be added 2 sites each from recN (10) and the Col A plasmid (11) and one site each from recQ (12), Col 1b (13), and Col E2 (14). These 19 sites have been used as the basis set for the calculation of the basepair frequencies and the various other quantities that can be derived from these.

Sequence Analysis

Wertman & Mount (15) have presented a qualitative analysis of the relation between sequence choice and function for the LexA sites. With the theory outlined above the statistical measures for the sequence choice can be quantitated and the correlation substantiated. The LexA sites are essentially symmetrical and we shall use both strands of the sequences giving a basis set of 38 sequences for the statistical analysis. The sites are listed in Table I along with their heterology index, λE, calculated from Eq. (2) and some binding constants and repression levels.

The base-pair frequencies are shown in Figure 1 together with the perceived consensus sequence and the sequence information calculated from Eq. (4). The sequence information provides a measure of the importance of the base-pair choice at the individual positions in the sites. Three base pairs in each halfsite stand out as being the most important; in the central region between these two main recognition elements there is a requirement for an alternating A-T sequence.



A	12	25	3	0	0	1	23	4	24	4	30	8	30	9	28	0	38	0	3	20
C	4	7	33	0	0	0	0	4	0	3	1	6	0	6	8	38	0	2	3	2
G	2	3	2	0	38	8	6	0	6	1	3	0	4	0	0	0	0	33	7	4
T	20	3	0	38	0	28	9	30	8	30	4	24	4	23	1	0	0	3	25	12
	T	A	C	T	G	T	A	T	A	T	A	T	A	T	A	C	A	G	T	A

Figure 1. The frequencies of base-pair occurrences and distribution of sequence information over the base-pair positions in the SOS boxes as listed in Table I. Every sequence have been counted twice, both as listed and the inversion. The consensus sequence deduced from the base pair frequencies listed is given on the bottom.

In Table I the contributions to λE from the two halves of each site are listed separately. In contrast to the situation for the recognition sites for the cyclic-AMP receptor protein (2) there does not seem to be any requirement for asymmetry in the sites; the two halfsite contributions to λE seem to be distributed evenly.

From Eq. (3) we expect a linear correlation between the heterology index and the binding strength of individual sites. This is shown to hold quite well in Figure 2. This figure includes both binding constants measured *in vitro* and binding constants estimated from *in vivo* repression data using Eq. (6). The slope of the correlation line gives a value of approximately 1.3 for the selection parameter λ; considering the heterogeneity in the sources of the experimental data and the statistical smallsample uncertainty in the estimates of λE as indicated by the error bars, this numerical estimate is of course very uncertain and would perhaps be better given as between 1.1 and 1.5.

TABLE I SOS boxes

site		sequence	λE_{l}^{a}	λE_{r}^{b}	λE^{c}	$\ln K^{d}$	$\ln IR^{e}$
recA		TACTGTATGAGCATACAGTA	3.2	3.5	6.7	20.0	2.4
uvrA		TACTGTATATTCATTGAGGT	0.0	7.9	7.9	17.2	
uvrB		AACTGTTTTTTTATCCAGTA	2.5	3.1	5.6	17.7	
sulA		TACTGTACATCCATACAGTA	1.9	4.3	6.2		
uvrD		ATCTGTATATATACCCAGCT	2.5	5.0	7.5		
mucAB		TACTGTATAAATAAACAGTT	1.9	1.4	3.3		
clo13		TACTGTGTATATATACAGTA	1.3	0.0	1.3		4.1
lexA	1	TGCTGTATATACTCACACGA	2.0	8.3	10.3	17.7*	2.1
lexA	2	AACTGTATATACACCCAGGG	0.5	6.6	7.1		
cle1	1	TGCTGTATATAAAACCAGTG	2.0	4.7	6.7	21.6*	
cle1	2	CAGTGGTTATATGTACAGTA	6.2	1.9	8.1		
Col1b		TACTGTATATGTATCCATAT	0.0	8.1	8.1		
ColA	1	TACTGTATATAAACACATGT	0.0	6.3	6.3	19.8*	
ColA	2	ACATGTGAATATATACAGTT	7.2	0.5	7.7		
ColE2		ATCTGTACATAAAACCAGTG	4.4	4.7	0.1		
umuDC	2	TACTGTATATAAAAACAGTA	0.0	2.0	2.0	22.3	
recN	1	TACTGTATATAAAACCAGTT	0.0	3.7	3.7		
recN	2	TACTGTACACAATAACAGTA	4.1	3.9	8.0		
recQ		GCCTGTTTTTATTT-CAGGC	5.3	5.2	10.5^{+}		

a Heterology index for the left half of the site

b Heterology index for the right half of the site

c Total heterology index for the whole site

d Natural logarithm of the binding constant expressed in M^{-1}; data and original references (when applicable) given by Bertrand-Burggraf *et al.*(22)

e Natural logarithm of the induction ratio [actually the induction ratio minus one; cf. Eq. (6)]; data from Wertman & Mount (15)

* The binding constant presumably refers to binding at both sites in the operon

+ The recQ sequence requires a different spacing between the two halves of the binding site; this will further increase the heterology index

The binding constants for three of the sets of tandem sites listed in Table I would fall close to the correlation line if the heterology index for the strongest site only is considered in each case. This is an indication that binding of LexA protein is given primarily by the affinity of the strongest site without any significant cooperative contribution from the weaker one in these cases.

In vivo Function

Let us consider the binding of LexA protein to the *recA* operator under repressed conditions *in vivo*. From the induction ratio of 11.5 (15) for this site and the *in vitro* binding constant $K\approx 5\cdot 10^{8}$ M^{-1} one finds from Eq. (6) that the concentration of free LexA in the cytoplasm would be approxi-

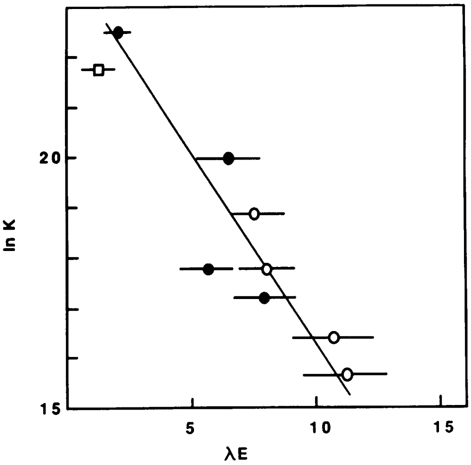


Figure 2. Correlation between the sequence heterology index λE and the logarithm of the binding constant for the LexA protein to single sites from the data listed in Table I. ● refer to in vitro binding constants and □ refer to binding constants estimated from the in vivo induction ratio relative to the binding at the recA site. The four data points indicated by o refer to single base-pair mutations of the recA operator with induction ratios given by Wertman & Mount (15). The error bars are determined from the small-sample uncertainty in the heterology index (1,2). The correlation line is a least-squares line minimizing deviations in λE.

mately $2 \cdot 10^{-8}$ M; this corresponds to about 12 free lexA dimers per cell which is a reasonable value. Actually, the binding constant is not necessarily the same under physiological conditions and this estimate of the free concentrations could be substantially off.

Furthermore, we can consider the induction ratio at the lexA operon containing two operator sites. Using $K_0R_f=10.5\exp(6.7/\lambda)$ from the recA data discussed above and the heterology indices $\lambda E_1=10.3$, $\lambda E_2=7.1$ (see Table 1) one finds from Eq. (5) the expected induction ratio:

$$1/P_0 = 1 + 10.5[\exp(-3.6/\lambda) + \exp(-0.4/\lambda) + w10.5\exp(-4/\lambda)] \qquad (9)$$

If $\lambda=1.3$ and $w=1$, this gives $1/P_0=14$ not too different from the observed induction ratio 8.8 (15), thus confirming the expectation that there is no significant cooperativity ($w \approx 1$) between the sites. For the two operator mutations pKW13 (changing λE_1 from 10.3 to 9) and pKW14 (changing λE_2 from 7.1 to 5.8) one finds in the same way $1/P_0=24$ and $1/P_0=36$, respectively.

This compares favourably with the observed induction ratios 18.5 and 29.8, respectively (15).

To estimate the amount of protein bound at pseudosites from Eq. (8) one can use K_sR_f=10.5 from the induction ratio of the operator site for *recA* since this site is about average in strength (Table I). Using I_{SEQ}=12, λ=1.3, and $\lambda^2\sigma^2$=17.5 from the sequence statistics one finds that $m_s \approx 40$. This is a reasonable estimate considering that there are between 100 and 1000 LexA protein molecules in the cell (16).

The expected number of random pseudosites in the genome that are equal to or stronger than an average LexA site can be estimated with the formula derived previously (1). One finds that the expected number of such strong pseudosites is around 5-10. Most of these, however, may not be accessible for binding.

ARG BOXES

Basis set

The arginine repressor controls several operons involved in arginine metabolism and Cunin *et al.*, (17) in a recent review have listed 9 recognition sequences. Most of these sites occur in pairs with a fixed spacing between them. Recently, a second ARG box has been identified in the *argR* operon with a shorter spacing (18). To these 10 binding sequences can be added two from *argD* (A. Boyen, personal communication) and three from the *argG* operon (F. Van Vliet, personal communication).

Experimental data in at least one case (19,20) indicate that the arginine repressor binds cooperatively to the two neighboring sites to turn off transcription initiation. Since the spacing between tandem sites is the same in all cases (except *argR*), it is reasonable to expect that binding is cooperative for all of them.

Sequence Analysis

We shall use the 15 sequences listed in Table II and their inversion giving a total of 30 symmetrized sites for statistical analysis. It is not known whether the third binding site in the *argG* operon is functional (N. Glansdorff *et al.*, personal communication); in this analysis it has been included in the basis set of binding sites but not ascribed any function. The base-pair frequencies are listed in Figure 3 which also depicts the sequence information as distributed over individual positions in the sites. Based on the sequence information one can identify four important base pairs in each

TABLE II ARG boxes

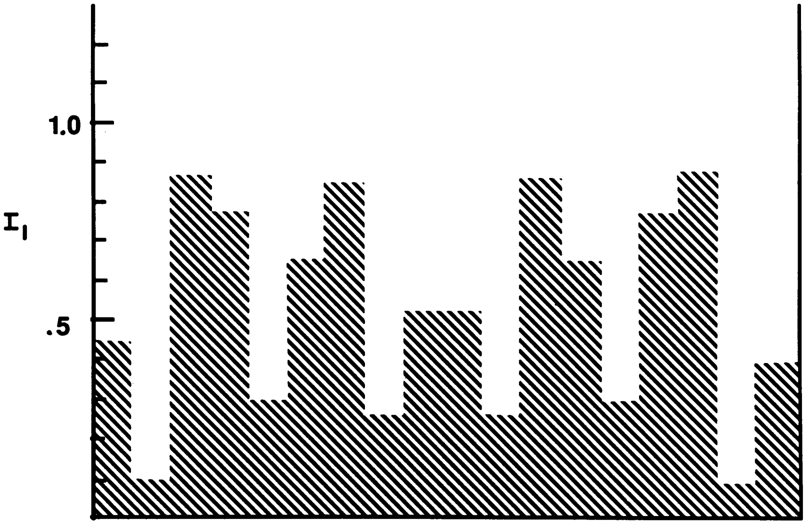
site	sequence	λE_{ℓ}^{a}	λE_{r}^{b}	λE^{c}	λE_{sum}^{d}	$\ln IR^{e}$
argF	1 AATGAATAATTACACATA	0.0	4.2	4.2	5.4	5.2
argF	2 AGTGAATTTTAATTCAAT	0.5	0.7	1.2		
argI	1 AATGAATAATCATCCATA	0.0	2.8	2.8	3.8	5.9
argI	2 ATTGAATTTTAATTCATT	0.9	0.1	1.0		
argECBH	1 TATCAATATTCATGCAGT	3.4	2.7	6.1	8.8	3.7,4.1
argECBH	2 TATGAATAAAAATACACT	0.2	2.5	2.7		
carAB	1 TGTGAATTAATATGCAAA	0.5	1.9	2.4	8.4	3.5
carAB	2 AGTGAGTGAATATTCTCT	3.1	2.9	6.0		
argR	1 TTTGCATAAAAATTCATC	1.7	2.7	4.4	15.2	2.2
argR	2 TATGCACAATAATGTTGT	4.0	6.8	10.8		
argD	1 AGTGATTTTTTATGCATA	2.9	1.1	4.0	12.3	3.0
argD	2 TGTGGTTATAATTTCACA	4.6	3.7	8.3		
argG	1 ATTAAATGAAAACTCATT	3.9	2.3	6.2	9.2	2.6
argG	2 TTTGCATAAAAATTCAGT	1.7	1.3	3.0		
argG	3 TGTGAATGAATATCCAGT	1.3	2.9	4.2		

a Heterology index for the left half of the sequence
b Heterology index for the right half of the sequence
c Total heterology index for the whole site
d Sum of the heterology indices for the two sites in the operon
e Natural logarithm of the induction ratio [actually the induction ratio minus one; cf. Eqs. (6) and (7)]; data from Cunin et al. (1986). The two numbers for argECBH refer to the expression of argE and argCBH respectively

halfsite; in the central region between these two recognition elements there is a requirement for A-T richness. The total sequence information (I_{SEQ} = 9.6) is somewhat lower than for the SOS boxes (I_{SEQ} = 12.1) and one would expect a larger presence of random pseudosites in the genome for the ARG boxes.

If the repression indeed is dominated by cooperative binding to two tandem sites, we expect from Eq. (7) that the sum of the heterology indices for the tandem sites will be linearly correlated with the logarithm of the induction ratio. This expectation is shown to hold quite well in Figure 4. The slope of the least-squares correlation line gives a value of approximately 2.1 for the selection parameter λ. Despite the numerical uncertainty, this value is significantly different from the result for LexA protein above (λ=1.3) and also from the results for the promotors (λ=1.0; ref.1) and the cyclic-AMP receptor protein binding sites (λ=0.7-0.8; ref. 2).

There is an equally strong correlation between the induction ratio and the heterology index of the strongest site in each operon as seen in Figure 5. This would be the expected result if binding at the weaker site is irrelevant or possibly anticooperative with binding at the stronger site [$w<1$ in



A	16	11	2	2	18	25	1	13	17	13	12	28	2	2	0	28	6	13
C	0	4	0	1	7	0	1	0	0	0	5	0	3	3	27	0	9	1
G	1	9	0	27	3	3	0	5	0	0	0	1	0	7	1	0	4	0
T	13	6	28	0	2	2	28	12	13	17	13	1	25	18	2	2	11	16
	A	A	T	G	A	A	T	A	A	T	T	A	T	T	C	A	T	T

Figure 3. Base-pair frequencies and distribution of sequence information over the base-pair positions in the ARG boxes as listed in Table II. Every sequence have been counted twice, both as listed and the inversion.

Eq. (5)] so that Eq. (6) is applicable. In this case the slope of the correlation line gives λ=0.9 which is similar to the estimates for the other systems considered.

Thus, judging from the statistical correlation between sequence and function, there are two possibilities: either binding to tandem sites is strongly cooperative as indicated by Figure 4, or it is dominated by the stronger site only as given by Figure 5. Obviously the statistical correlation is sufficiently good in both cases so that neither can be excluded; note the large statistical uncertainties given by the error bars in Figures 4 and 5 indicating that any small difference in correlation carries little weight in determining which case is the most likely. However, the experimental evidence quoted above indicates that binding is cooperative; also the fact that the spacing between tandem sites is (almost) conserved -- in contrast to the situation for the SOS boxes -- supports the notion that cooperative binding is important for activity.

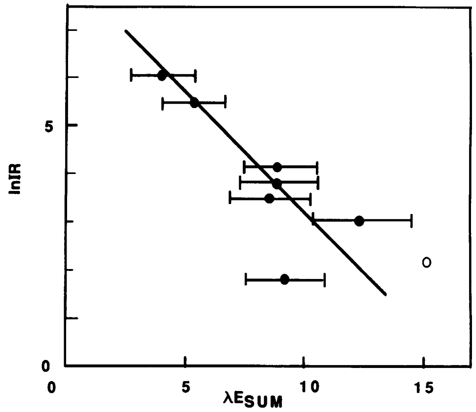


Figure 4. Correlation between the sum of the heterology indices for the tandem ARG boxes and the logarithm of the *in vivo* induction ratio of the relevant operon from the data listed in Table II.

Thus it seems reasonable that the binding is strongly cooperative and that the correlation displayed in Figure 5 is a consequence only of the fact that the strongest site also provides the dominant contribution to the cooperative binding.

In the analysis of the correlation, the result for *arg*R(o in Figure 4) has not been included although it falls close to the correlation line. The

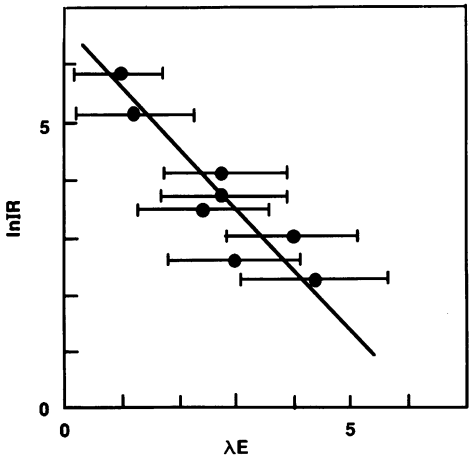


Figure 5. Correlation between the heterology index of the strongest site and the logarithm of the *in vivo* induction ratio of each *arg* operon from the data listed in Table II.

agreement could be fortuitous since the cooperativity of binding at the tandem argR sites separated by only two base pairs is expected to be different from all the other cases where the spacing is three base pairs. The agreement on the other hand indicates that the cooperativity is not much different for tandem binding to the ARG boxes in the argR operon.

In vivo Function

Let us consider the function of the two ARG boxes regulating the argECBH operons. These have the heterology indices $\lambda E_1=6.1$ and $\lambda E_2=2.7$ (Table II) respectively. The average induction ratio for the argE and the argCBH operons is $1/P_0=50$. When the spacing between the two ARG boxes is reduced by one, the induction ratio is reduced to $1/P_0=2$ (19,20). This reduction could be achieved in several ways. The simplest interpretation is that the spacing change reduces the cooperativity factor w. The experimental data indicate that cooperative binding still dominates (see below) such that Eq. (7) is applicable also for the spacing mutant; a reduction in $1/P_0$ from 50 to 2 corresponds to a reduction in w by a factor ~ 50. In this case nothing can be said about the values of w or K_0R_f separately.

However, limit values for w and K_0R_f can be estimated from the assumption that simultaneous binding to the two sites becomes sterically impossible when the spacing is reduced; then one finds by setting w=0 in Eq. (5)

$$2 = 1 + K_0R_f[\exp(-6.1/\lambda)+\exp(-2.7/\lambda)] \qquad (10)$$

If $\lambda=2.1$ as determined from the correlation line in Figure 4 this gives $K_0R_f=3.0$. Using this result when Eq. (5) is applied to the wild-type induction ratio, one finds

$$50 = 1 + K_0R_f[\exp(-6.1/\lambda)\exp(-2.7/\lambda)+wK_0R_f\exp(-8.8/\lambda)] \qquad (11)$$

which gives the cooperativity factor w=350.

Since it is not likely that cooperative binding is sterically impossible when the spacing is reduced, these estimates represent -- within the statistical uncertainties -- an upper limit for K_0R_f and a lower limit for w. In fact, the experimental data indicate that cooperative binding still dominates; otherwise the spacing reduction would be expected to influence the expression of argCBH differently from argE since the repressor would be bound predominantly at the second site which would block initiation at the main argE promotor but probably not hinder argCBH expression. The experimental data (19,20) show that the repressibility of both is changed in approximately the same way indicating that both promotor sites are blocked similarly and that the repressor binds simultaneously to both ARG boxes. Also the fact that the repressor binds simultaneously to the tandem sites in the argR operon where the spacing is shorter in the wildtype (18) supports the view that cooperative biding dominates in all these cases.

A cooperative binding could imply that the base-pair choices in the spacer region are important. This is corroborated by the fact that the spacers including the neighbouring base pairs in the binding site on each side consist of five A•T or T•A base pairs in various configurations. The exception is the short spacer in the argR operon; including the two base pairs at each side, it has the sequence CTGT. It is possible that the unusual base-pair combination in this spacer can partially compensate for it being shorter than the others.

The extent of pseudosite binding can be estimated from these results using Eq. (8). Assuming first that most of the pseudosites will be occurring singly (the probability for two sites occurring next to each other is much smaller), the number of arginine repressors bound at single pseudosites is found to be $m_S \approx 35 K_0 R_f$ using $\lambda E_{SEQ}=4.5$, $I_{SEQ}=9.6$, $\lambda^2 \sigma^2=14.4$ from the pase-pair statistics and $\lambda=2.1$ from the correlation line in Figure 4. If $K_0 R_f \approx 3$ as estimated above this gives $m_s \approx 100$. Since there are no more than 40 to 200 arginine repressors in the cell (21), the upper limit of the estimate is clearly impossible. This could be taken as a further indication that cooperativity is dominant also in the spacing mutant of argECBH which was used for the upper-limit estimate of $K_0 R_f$ above. However, the calculation of m_s is very uncertain as it relies on several inadequately known numbers.

The extent of binding to random tandem sites can also be estimated from Eq. (8) using values for I_{seq} and σ^2 that are double those estimated for single sites. Furthermore, the factor $K_s R_f$ in Eq. (8) should be replaced by the corresponding saturation level of an average specific tandem site. In this way one finds $m_s \approx 10$. Although the specific binding is determined by tandem sites, the pseudosite competition and therefore also the functional specificity may be dominated by single sites.

The number of strong tandem pseudosites in the genome can be estimated to be ~ 0.005 using the formula derived previously (1). This small number is of course due to the low probability of finding two sites next to each other; or equivalently, due to the fact that the sequence information for tandem sites is twice that of the single sites. This small number also implies that the estimated pseudosite competition is dominated by very weak sites.

CONCLUSIONS AND DISCUSSION

The statistical analysis of the recognition sequences provides information on the specificity and function of the regulatory systems in question. The two repressor systems analyzed in this communication as well as the activator system analyzed previously (2) display significant differences in this regard.

The functional specificity as determined by the estimated pseudosite competition is higher for the two repressor systems than for the activator (2). While the total pseudosite competition is determined mostly by the sequence information [cf. Eq. (8)], the value of λ will determine what kind of sites that dominate this competition. When λ is smaller than one, the competition is dominated by sites that are similar to the specific ones in binding strength; this is the case expected for the cyclic AMP receptor protein (2). In contrast, the two repressor systems described above with larger λ-values are expected to have a pseudosite competition that is dominated by weaker sites. In this way, the random occurrence in the genome of a large number of strong repressor sites is avoided without a substantial increase in functional specificity. This situation is most pronounced for the ARG boxes which have the largest λ-value.

The sequence analysis supports the proposition that tandem ARG boxes are strongly cooperative while the SOS boxes are not. Furthermore, the individual base-pair contributions to the binding strength are smaller for the ARG boxes than for the SOS boxes. Thus, it can be expected that the function of the ARG boxes is less sensitive to point mutations. In some sense this is partially balanced by the fact that in the tandem sites there are more possibilities for point mutations to occur. Moreover, the relatively small differences in binding strength expected after changes in the sequence of an ARG box are probably related to the expected weak binding to a single cognate (consensus) ARG box.

The tandem ARG boxes is the only case studied so far that appears to be overspecified (4) in that their probability for random occurrence in the genome is much smaller than one. This might be necessary in this case in order to keep the competition from single pseudosites at a reasonable level.

If the arginine repressor exclusively uses cooperative tandem sites as regulatory units, it seems very likely that induction is based on a reduction of the cooperativity parameter w rather than on a change in the DNA binding constants of the repressor in the presence of arginine. This is reasonable since in this picture the main contribution to the binding saturation is given by the cooperativity while single-site binding is very weak. Thus, the large λ-value and the expected weak single-site binding may both be a consequence of a different mode of induction of the arginine regulon.

The LexA protein is expected to have much stronger binding to its individual recognition sites than the arginine repressor. The pseudosites for the LexA protein is expected to be much smaller in number but also stronger in binding. A substantial binding for the arginine repressor on the other hand can be achieved only by cooperative binding at tandem sites. The pseudosite binding for the arginine repressor estimated above is dependent on a very low binding saturation at a very large number of weak pseudosites.

The statistical analysis carries a large uncertainty due to the small samples of sites considered. The experimental data are quite uncertain as well, particularly the _in vivo_ estimates where gene expression could be influenced by many factors other than repressor binding. However, also binding constants measured _in vitro_ have been measured in different ways and under different conditions and are therefore not always quantitatively comparable. Nevertheless, the statistical correlations between sequence and function displayed in Figures 2 and 4 demonstrate that the selection theory developed previously holds up well. Based on the theoretical relations and some experimental data it is possible to predict quantitatively the functional activity of particular recognition sequences. As demonstrated here and previously, the theory provides a useful framework within which specificity, sequence choice and function can be discussed.

ACKNOWLEDGEMENTS

I am very grateful to Drs. N. Glansdorff, A. Boyen, F. Van Vliet, R. Cunin and D. Charlier (University of Brussels) for communicating their unpublished sequence data and for useful comments. This work has been supported by the Swedish Natural Science Research Council.

REFERENCES

1. Berg, O.G. & von Hippel, P.H. (1987) J. Mol. Biol. 193, 723-750.
2. Berg, O.G. & von Hippel, P.H. (1988) J. Mol. Biol., in press.
3. Schneider, T.D., Stormo, G.D., Gold L. & Ehrenfeucht, A. (1986) J.Mol. Biol. 188, 415-431.
4. von Hippel, P.H. (1979) In: Biological Regulation and Development (Goldberger, R.F., ed), Vol 1, Plenum, New York, pp. 279-347.

5. von Hippel, P.H. & Berg, O.G. (1986) Proc. Nat. Acad. Sci. USA 83, 1608-1612.
6. Little, J.W. & Mount, D.W. (1982) Cell 29, 11-22.
7. Walker, G.C. (1984) Microbiol. Rev. 48, 60-93.
8. Peterson, K.R. & Mount, D.W. (1987) J. Mol. Biol. 193, 27-40.
9. Kitagawa, Y., Akaboshi, E., Shinagawa, H., Horii, T., Ogawa, H. & Kato, T. (1985) Proc. Natl. Acad. Sci. USA 82, 4336-4340.
10. Rostas, K. Morton, S.J., Picksley, S.M. & Lloyd, R.G. (1987) Nucl. Acids Res. 15, 5041-5049.
11. Lloubes, R., Baty, D. & Lazdunski, C. (1986) Nucl. Acids Res. 14, 2621-2636.
12. Irino, N., Nakayama, K. & Nakayama, H. (1986) Mol. Gen. Genet. 205, 298-304.
13. Mankovich, J.A., Lai, P.-H., Gokul, N. & Konisky, J. (1984) J. Biol. chem. 259, 8764-8768.
14. Cole, S.T., Saint-Joanis, B. & Pugsley, A.P. (1985) Mol. Gen. Genet. 198, 465-472.
15. Wertman, K.F. & Mount D.W. (1985) J. Bact. 163, 376-384.
16. Brent, R. & Ptashne, M. (1984) Nature 312, 612-615.
17. Cunin, R., Glansdorff, N., Pierard, A. & Stalon, V. (1986) Microbiol. Rev. 50, 314-352.
18. Lim, D., Oppenheim, J.D., Eckhardt, T. & Maas, W.K. (1987) Proc. Natl. Acad. Sci. USA 84, 6697-6701.
19. Piette, J., Cunin, R., Boyen, A., Charlier, D., Crabeel, M., Van Vliet, F., Glansdorff, N., Squires, C. & Squires, C.L. (1982) Nucl. Acids Res. 10, 8031-8048.
20. Cunin, R., Eckhardt, T., Piette, J., Boyen, A., Pierard, A. & Glansdorff, N. (1983) Nucl. Acids Res. 11, 5007-5019.
21. Lissens, W., Cunin, R., Kelker, N., Glansdorff, N. & Pierard, A. (1980) J. Bact. 141, 58-66.