*J. Mol. Biol.* (1987) **193**, 723–750

# Selection of DNA Binding Sites by Regulatory Proteins

## Statistical-mechanical Theory and Application to Operators and Promoters

**Otto G. Berg† and Peter H. von Hippel**

*Institute of Molecular Biology*
*University of Oregon*
*Eugene, Oregon 97403, U.S.A.*

*(Received 7 May 1986, and in revised form 29 September 1986)*

We present a statistical-mechanical selection theory for the sequence analysis of a set of specific DNA regulatory sites that makes it possible to predict the relationship between individual base-pair choices in the site and specific activity (affinity). The theory is based on the assumption that specific DNA sequences have been selected to conform to some requirement for protein binding (or activity), and that all sequences that can fulfil this requirement are equally likely to occur. In most cases, the number of specific DNA sequences that are known for a certain DNA-binding protein is very small, and we discuss in detail the small-sample uncertainties that this leads to. When applied to the binding sites for *cro* repressor in phage lambda, the theory can predict, from the sequence statistics alone, their rank order binding affinities in reasonable agreement with measured values. However, the statistical uncertainty generated by such a small sample (only 6 sites known) limits the result to order-of-magnitude comparisons. When applied to the much larger sample of *Escherichia coli* promoter sequences, the theory predicts the correlation between *in vitro* activity ($k_2 K_B$ values) and homology score (closeness to the consensus sequence) observed by Mulligan *et al.* (1984). The analysis of base-pair frequencies in the promoter sample is consistent with the assumption that base-pairs at different positions in the sites contribute independently to the specific activity, except in a few marginal cases that are discussed. When the promoter sites are ordered according to predicted activities, they seem to conform to the Gaussian distribution that results from a requirement for maximal sequence variability within the constraint of providing a certain average activity. The theory allows us to compare the number of specific sites with a certain activity to the number that would be expected from random occurrence in the genome. While strong promoters are "overspecified", in the sense that their probability of random occurrence is very low, random sequences with weak promoter-like properties are expected to occur in very large numbers. This leads to the conclusion that functional specificity is based on other properties in addition to primary sequence recognition; some possibilities are discussed. Finally, we show that the sequence information, as defined by Schneider *et al.* (1986), can be used directly (at least in the case of equilibrium binding sites) to estimate the number of protein molecules that are specifically bound at random "pseudosites" in the genome. This provides the connection between base-pair sequence statistics and functional *in vivo* specificity as defined by von Hippel & Berg (1986).

### 1. Introduction

Genome-regulatory proteins recognize and bind to specific DNA sites among a vast excess of structurally similar non-specific sites (e.g. repressor–operator or RNA polymerase–promoter interactions). Such binding selection derives from specific interactions between the active site of the protein and the base-pairs in the DNA binding sequence. In a recent paper (von Hippel & Berg, 1986) we discussed the molecular origins of this specificity, and also explored the requirements for specific binding site selection in the living cell. These requirements stem from the competition for

† Present address: Department of Molecular Biology, The Biomedical Center, Box 590, S-75124 Uppsala, Sweden.



0022-2836/87/040723-28 $03.00/0

protein by non-specific sites that are close to the specific ones in sequence and/or in binding affinity (von Hippel, 1979). To quantify the extent of this non-specific competition, and thereby the magnitude of the effective binding selection process, we need to know the magnitude of the reduction in specific binding affinity that accompanies the insertion of each of the "wrong" base-pairs that distinguish a particular non-specific site from the specific site. One way of doing this is to isolate or create sites with various degrees of homology to the specific sites and to measure the resulting binding constants (e.g. see Jobe *et al.*, 1974; Mossing & Record, 1985).

In the absence of exhaustive binding data, one can still make some inferences about the relative importance of specific base-pair interactions at different positions within the binding site. Positions in the site where the base-pairs vary greatly between the specific sequences can be expected to contribute little to the binding specificity, while conserved or nearly conserved base-pairs doubtless contribute a great deal. Thus, Schneider *et al.* (1986) have used information theory to quantify the importance of particular base-pairs, based on their variability in the specific sequences. However, this measure of information is not directly related to the relative binding affinities, and thus cannot tell us anything quantitative about specific interactions. Mulligan *et al.* (1984) have found a correlation between the activity of promoter sites and a "homology score" that measures the closeness of a particular sequence to the perceived consensus promoter sequence. This homology measure weighs the importance of each "wrong" base-pair (relative to the consensus sequence) against the observed variability among all promoters. Mulligan *et al.* (1984) and Mulligan & McClure (1986) also used this homology score as a basis for a computer search for promoters in known DNA sequences. Similar search algorithms, using somewhat different weighting schemes, have been presented and applied by others (Harr *et al.*, 1983; Staden, 1984). Recently, a more elaborate pattern-recognition method has been designed and used to analyze *Escherichia coli* promoter sequences (Galas *et al.*, 1985).

While the results of the present paper will be relevant to the design of search algorithms and to the interpretation of sequence analyses, our main focus will be on the relationship between sequence variability within a set of specific sites and the interaction free energy contributed by each base-pair in the site. The connection between sequence variability and binding affinity derives from an evolutionary selection constraint. That is, specific binding sequences can be assumed to be selected to show binding affinities in some useful range. Fortunately, we do not need to understand the exact nature of this selection constraint in order to establish the required theoretical connection between sequence variability and binding free energy. In the next section we explore the consequences of some different selection constraints and derive the desired relations between the sequence variability in the set of possible binding sequences and the corresponding interaction free energies. In the third section the theory is first applied to the binding sites for the *cro* repressor of lambda phage and is then extended and applied to the *E. coli* and coliphage promoter sequences studied by Mulligan *et al.* (1984). Our theory suggests a more general homology measure that is directly related to binding and activity; this measure more closely resembles the statistical weighting scheme used by Harr *et al.* (1983). The significance of the information measures used by Schneider *et al.* (1986) is also discussed in light of the present results, and it is shown how they relate to the functional specificity requirements in the living cell. In the fourth section we explore the consequences of the theory for the evolutionary selection of binding interactions. It is obvious that specificity is not maximized in evolution. Instead we argue that evolution *minimizes the maximum loss* of specificity, in the sense that specificity will tend towards a situation where mutational errors have relatively small effects.

The statistical-mechanical selection theory provides a physical basis for the analysis and interpretation of sequence data. It enables us to quantify the expected specificity for any sequence and sets that in relation to the requirements for its biological function. A statistical analysis of this sort requires a fairly large sample of specific sites to provide a reasonable predictive accuracy. Thus, all quantitative results are compared to the expected statistical uncertainties. However, the general results of the theory are not dependent on actual sequence analysis; thus, they provide a framework within which the relationships between sequence variability, specificity and function can be understood.

## 2. Statistical-mechanical Ensembles that Describe the Sequence Variability of Specific Binding Sites

Protein–DNA recognition is based primarily on the DNA-sequence-dependent hydrogen bond donor and acceptor patterns exposed in the grooves of the double helix. These patterns must be more or less complementary to similar patterns in the binding site of the protein. In particular cases these interactions have been identified physically (e.g. by X-ray crystallography), but in general the structure of the recognition site on the protein is not known. One can also gain information about the importance of base-pair interactions in a recognition sequence by studying the effects on recognition of the modification or substitution of individual base-pairs. In the absence of detailed laboratory studies of such effects, the sequence analysis of naturally occurring recognition sites can provide similar information. This follows because we assume that in the course of evolution nature has carried out analogous experiments, testing base-pair substitutions and accepting and rejecting sequences on the

basis of their properties as recognition sites. In recent years the number of known sequences has been increasing sharply, while the physical characterization of their functional properties (binding, activity, etc.) has lagged far behind. Thus, it becomes particularly important to derive as much information as possible from the sequence analysis itself.

### (a) *Consensus sequences*

A certain sequence-specific ("recognizer") protein can normally recognize and bind to DNA sites that vary somewhat in sequence, so that a base-pair at any particular position may differ from site to site. In general such sites share common features, consensus base-pairs that "almost always" appear at the same position in every site. These consensus base-pairs then form a distinct pattern that can help the biologist to identify previously unknown recognition sites in other DNA sequences. Qualitatively, it can also be argued that sites that differ more from the perceived consensus sequence are weaker recognition sites, and this also appears to be so in cases that have been tested (e.g. for operators or promoters). In principle, the consensus sequence could be defined as the sequence that, at every position in the site, carries the base-pair most often found at this position in the set of all naturally occurring sites that have been sequenced. Obviously, with such an extended definition of consensus, not all consensus base-pairs turn out to be equally significant. Some questions that naturally arise are: (1) how should one quantify the significance of a consensus base-pair and the importance of deviations from it? (2) What can be deduced from the sequence data about the recognition mechanisms and the functional properties of particular sequences?

Various statistical measures have been applied in approaching the first question. For example, Schneider *et al.* (1986) have used information theory to analyze some sets of operator sites and have assigned a measure that quantifies the importance of the base-pair choice at each position in a particular kind of recognition site. This measure of information is determined from the probability that the observed base-pair utilization frequency has appeared at random. Thus, the information-theoretic sequence analysis can also be used to estimate the probability that a certain kind of recognition sequence will occur at random in the genome. However, a statistical measure of this sort cannot tell us anything quantitative about the recognition efficiency (binding or activity) of a particular base-pair sequence. Mulligan *et al.* (1984) devised a measure for the importance of any particular base-pair choice at individual positions in promoter sites, based on its frequency of occurrence in the set of identified promoter sites found in *E. coli*. Although it is not obvious why this *particular* measure should be chosen, Mulligan *et al.* (1984) used it to quantify a homology score (defined as the degree of closeness of a given promoter to the consensus sequence) and were able to demonstrate a correlation between this core and the activity of various promoter sequences.

What has been missing in most sequence studies of this sort is an *a priori* coupling between sequence choice and functional properties. In order to use the sequence analysis of the naturally occurring recognition sites to predict the recognition efficiency of any particular sequence, one must know what constraints the sites were chosen to satisfy. In the course of evolution, only sequences with binding affinity (or activity) in some useful range would be selected as specific recognition sites. These evolutionary selection constraints provide the necessary relationship between sequence choice and functional properties. Rather than just guessing what this relationship might be, we shall proceed by assuming that *some* selection constraint is operating and then consider all possible sequences that could satisfy it.

### (b) *Selection model*

For simplicity of discussion, let us first consider a set of sites that have been chosen on the basis of affinity for a particular sequence-specific protein (e.g. operator sites). Differences in base-pair choice at certain positions in individual sites can have several causes. (1) Certain sites may require a different binding affinity depending on their functional role in the genome. (2) Some base-pair choices may be neutral with respect to binding affinity; or if they do matter, the binding affinity could be compensated by appropriate choices at other positions in the site. (3) Some base-pair choices may be required for regulation (e.g. binding of effector molecules) rather than for binding of the recognizer protein under consideration. The effect of a requirement of this type is difficult to predict without knowing its exact nature and will not be taken into account in the derivations below. However, the results of the analysis make it possible to identify and discuss the effects of some such requirements.

Thus, we shall consider the effects of the binding requirements for one particular recognizer protein on base-pair variability (or base-pair conservation) within the set of binding sites. To be able to do this we shall assume that the binding free energies for all possible sequences are known and then derive the most probable base-pair utilization frequencies that ensue. This is analogous to a statistical-mechanical approach in which it is assumed that the energy levels of a particular system are known; a distribution of level occupancies can then be calculated. In the Appendix these distributions are calculated from first principles. In this section we shall pursue the statistical-mechanical analogy.

The basic assumptions we make are: (1) individual binding sequences are selected to have a value of binding affinity for the recognizer protein in some useful range. Depending on the functional role of the protein–DNA interaction at issue, this range

may well vary between individual sites. (2) The number of sequences in such an affinity range that could possibly be used is large. If selection is only on the basis of affinity, "neutral sequence drift" within this selection constraint will ensure that all possible sequences are equiprobable. (3) Each possible base-pair $B$ ($B = 0, 1, 2, 3$, where e.g. $0 = \mathrm{A \cdot T}$, $1 = \mathrm{T \cdot A}$, $2 = \mathrm{G \cdot C}$, and $3 = \mathrm{C \cdot G}$) at position $l$ ($l = 1, 2, \ldots, s$, where $s$ is the site size) in a binding site contributes a certain amount $\varepsilon_{lB}kT$† to the binding free energy at that site. These individual base-pair contributions are assumed to be independent and therefore additive. The strongest binder, the cognate base-pair designated $B = 0$ at each position, is considered to define the ground-state level with $\varepsilon_{l0} = 0$. Thus, $\varepsilon_{lB}$ are dimensionless positive numbers that express the decrease in (favorable) binding free energy (in units of $kT$) that results when the cognate base-pair at position $l$ is replaced by base-pair $B$; this will be referred to in the following discussion as the local (per base-pair) *discrimination energy*.

The total discrimination energy for a particular sequence is given by the sum of the local contributions from the individual base-pairs. In this way the binding affinity for all sequences are measured relative to the best binding (cognate) sequence. (In principle any particular base-pair sequence could be chosen as the standard to which other sequences are compared; then $\varepsilon_{lB}$ could be either positive or negative.)

(i) *Sites selected with the same binding affinity*

Let us consider the potential binding sites as the set of all possible sequences that have binding affinity in some limited range around some fixed required value. All such sequences must have discrimination energy in some limited range $\Delta E$ around a required level $E$. Thus, in each potential site, the local contribution $\varepsilon_{lB}$ from every position $l$ must sum to $E$. In the set of all potential sites, what is the frequency with which a certain base-pair $B$ appears at a certain position in a site? This question can be answered by counting all possible sequence combinations that provide the required discrimination energy $E$ (see the Appendix). However, a completely equivalent question is frequently asked in statistical mechanics, where one seeks to describe the probability of energy-level occupancy given that the total energy should sum to a given value (e.g. see Gurney, 1949). Thus, a potential site can be considered as the realization of a statistical-mechanical system of $s$ independent particles and a given energy $E$. Choosing base-pair $B$ at position $l$ in a sequence corresponds to putting particle $l$ into energy level $\varepsilon_{lB}$ in the corresponding statistical-mechanical system. (Thus sequences are chosen according to a microcanonical ensemble.) To start with we shall assume that base-pairs are chosen with equal *a priori* probabilities, i.e. that they are equally common in the genome. (This assumption is removed in the Appendix.) Then, in analogy with the probability distribution over single-particle levels, the probability $f_{lB}$ of choosing base-pairs $B$ at position $l$ is proportional to the usual Boltzmann factor $\exp(-\lambda\varepsilon_{lB})$:

$$f_{lB}(E) = \exp(-\lambda\varepsilon_{lB})/4q_l; \qquad B = 0, 1, 2, 3 \quad \text{and} \quad l = 1, 2, \ldots, s, \quad (1)$$

where:

$$q_l = [1 + \exp(-\lambda\varepsilon_{l1}) + \exp(-\lambda\varepsilon_{l2}) + \exp(-\lambda\varepsilon_{l3})]/4 \quad (2)$$

is the partition function that is chosen to ensure that the base-pair probabilities in equation (1) sum to unity at each position $l$. The coupling factor $\lambda$ is a dimensionless number, which has to be chosen so that the distribution satisfies the selection constraint, i.e. so that the discrimination energy $E$ has the assumed value. In a sense, $\lambda$ compensates for the fact that, even if base-pairs contribute independently to the binding affinity, their frequency of occurrence cannot be totally independent since their contributions in each site must add up to the assumed value of $E$.

In the combinatorical derivation of statistical-mechanical energy distributions (e.g. see Gurney, 1949), a statistical parameter corresponding to $\lambda$ appears as in equation (1) in order to satisfy the constraints on overall energy. When it is required that the relations agree with classical thermodynamics, this parameter can be identified with the absolute temperature of the system as $\lambda = 1/kT$. Obviously, in the case of sequence selection described here, we are not concerned with a thermodynamical system. Thus, $\lambda$ has a less obvious physical interpretation, though it serves, in the same sense as $kT$, as a proportionality factor to relate populations of base-pair choices to binding free energies. In the Appendix we show that $\lambda$ is determined by the density of potential sites, i.e. by the number of possible sequence combinations that have the required discrimination energy $E$.

Since the free energy contributions of individual base-pairs to the binding affinity are assumed to be additive, $E$ can be calculated as the average over the whole set:

$$E = \sum_{l=1}^{s} \sum_{B=1}^{3} \varepsilon_{lB} f_{lB}. \quad (3)$$

Inserting $f_{lB}$ from equation (1), this gives an implicit relation from which $\lambda$ can be calculated. Thus, the base-pair utilization frequences $f_{lB}$ of equation (1) depend implicitly on $E$ through their dependence on $\lambda(E)$ *via* equation (3).

In the particularly simple case where all local discrimination energies are the same ($\varepsilon_{lB} = \varepsilon$), from equations (1), (2) and (3) one finds for a site comprising $s$ base-pairs that:

$$\lambda(E) = \ln(3s\varepsilon/E - 3)/\varepsilon. \quad (4)$$

While the selection parameter $\lambda$ in principle can

† $kT$, the product of the Boltzmann constant and the absolute temperature.

have any positive value, for practical reasons it seems that $\lambda$ varies between $\sim 0{\cdot}5$ and $\sim 1{\cdot}5$ at most.

(ii) *Sites selected with a distribution of binding affinities*

In general a set of sites will include all known binding sequences for a given protein. These sites may exhibit a wide variety of affinities for the protein, depending on their functional role in the genome. A set of $n_s$ binding sites will then have some distribution $g(E)$ of discrimination energies so that $n_s g(E)\Delta E$ is the number of sequences that have discrimination energies in the range $\Delta E$ around $E$. The observed base-pair utilization frequencies $f_{lB}^{\text{obs}}$ in this whole set will be an average over the expected base-pair frequencies in each affinity range:

$$f_{lB}^{\text{obs}} = \int f_{lB}(E)g(E)\mathrm{d}E. \tag{5}$$

As seen in equation (4), $\lambda(E)$ is relatively insensitive to changes in $E$. Thus, to a first-order approximation in variations of $\lambda$, equation (5) gives:

$$f_{lB}^{\text{obs}} = f_{lB}(\langle E\rangle_{\text{seq}}), \tag{6}$$

and, while equation (1) still holds, $E$ has been replaced by its average value over the whole set of sites:

$$\langle E\rangle_{\text{seq}} = \int Eg(E)\mathrm{d}E. \tag{7}$$

Again, from the assumed additivity of all individual base-pair contributions, as in equation (3) the selection energy $\langle E\rangle_{\text{seq}}$ is equal to the average over the base-pair utilization frequencies:

$$\langle E\rangle_{\text{seq}} = \sum_{l=1}^{s}\sum_{B=1}^{3} \varepsilon_{lB} f_{lB}^{\text{obs}}. \tag{8}$$

This expression consequently determines the parameter $\lambda(\langle E\rangle_{\text{seq}})$ if the individual $\varepsilon_{lB}$ terms are known. It should be stressed that $\lambda$ is thus a quantity determined from the properties of the whole set of sites, and (at least to a first-order approximation) does not vary from site to site.

Thus, the expected base-pair utilization frequencies are very insensitive to variations in the required discrimination energy, and are determined primarily from its average. Consequently every site in the set gives approximately the same contribution to the base-pair frequencies, regardless of its exact discrimination energy. It therefore follows that the base-pair frequencies from the whole set of binding sites can be analyzed properly and not just from sites with affinities in some limited range.

The statistical sequence analysis of a set of binding sites will provide the base-pair utilization frequencies, $f_{lB}^{\text{obs}}$. With these we can calculate the local discrimination energies, $\varepsilon_{lB}$, *via* equation (1) in the form:

$$\lambda\varepsilon_{lB}^{\text{obs}} = \ln\,(f_{l0}^{\text{obs}}/f_{lB}^{\text{obs}}) \tag{9}$$

and their average over all possible substitutions:

$$\lambda\bar{\varepsilon} = \frac{1}{3s}\sum_{l=1}^{s}\sum_{B=0}^{3} \lambda\varepsilon_{lB}^{\text{obs}}. \tag{10}$$

With this definition, $3s\bar{\varepsilon}/4$ corresponds to the average discrimination energy for a random sequence. The parameter $\lambda$, in principle determined by equation (8), remains undetermined from the sequence analysis unless real binding free energies are known for at least some sites. When all local discrimination energies $\varepsilon_{lB}$ are known from equation (9), one can calculate the actual discrimination energy:

$$E(\{B_l\}) = \sum_{l=1}^{s} \varepsilon_{lB_l} = \frac{1}{\lambda}\sum_{l=1}^{s} \ln\,(f_{l0}^{\text{obs}}/f_{lB_l}^{\text{obs}}) \tag{11}$$

for any sequence $\{B_l\}_{l=1}^{s}$. This will also provide information on the form of the required discrimination energy distribution $g(E)$ for the set of specific sites studied.

(iii) *Sequence information*

From the observed base-pair frequencies, Schneider *et al.* (1986) defined and calculated the information contained in a given set of sequences as:

$$I_{\text{seq}} = \sum_{l=1}^{s}\sum_{B=0}^{3} f_{lB}^{\text{obs}} \ln\,[f_{lB}^{\text{obs}}/p^{\circ}(B)], \tag{12a}$$

where $p^{\circ}(B)$ is the *a priori* probability of the occurrence of base-pair $B$. When all base-pairs are equally common in the genome, $p^{\circ}(B) = 1/4$, and one finds, using equations (8) and (9) that:

$$I_{\text{seq}} = -\lambda\langle E\rangle_{\text{seq}} - \sum_{l=1}^{s} \ln q_l. \tag{12b}$$

Thus, the sequence information is directly related to the average discrimination. In fact, in the statistical-mechanical analogy the negative of the sequence information serves as the "selection entropy" (see below). The connection of this selection entropy with the thermodynamic entropy, $S$, becomes even more clear when it is observed from equation (12b) that $(\mathrm{d}I_{\text{seq}}/\mathrm{d}\langle E\rangle_{\text{seq}} = -\lambda$, in analogy with the thermodynamic relation $(\mathrm{d}S/\mathrm{d}E) = 1/T$, in which $T$ is the absolute temperature.

In the Appendix the probability of random occurrence of a site with discrimination energy below some cut-off energy $E$ is calculated. From equations (A16) and (12b), one finds that:

$$P_s(\langle E\rangle_{\text{seq}}) = \frac{\exp\,(-I_{\text{seq}})}{[6\pi s(1-\bar{\varepsilon}/2\langle E\rangle_{\text{seq}})]^{1/2}(1-\langle E\rangle_{\text{seq}}/s\bar{\varepsilon})} \tag{13}$$

expresses the probability of random occurrence of a site with discrimination energy less than the average $\langle E\rangle_{\text{seq}}$ for the sites in the set studied. All the quantities required ($\lambda\langle E\rangle_{\text{seq}}$, $\lambda\bar{\varepsilon}$, and $I_{\text{seq}}$) to calculate this probability are determined from the sequence data *via* equations (8), (9), (10) and (12a). While the expression is dominated by $I_{\text{seq}}$ (and this is the primary physical meaning of the sequence

information) the correction factor is also significant. Equation (13) makes it possible to estimate the number of randomly occurring binding sites in the genome.

When the local discrimination energies ($\varepsilon_{lB}$) are known we can also calculate the average binding constant, $K_R$, for a random site. Since the random probability for any particular sequence $\{B_l\}$ of length $s$ is $4^{-s}$ and its binding constant is a factor $\prod_{l=1}^{s} \exp(-\varepsilon_{lB_l})$ smaller than that of the consensus sequence ($K_O$), one finds:

$$K_R = K_O 4^{-s} \sum_{\{B_l\}} \prod_{l=1}^{s} \exp(-\varepsilon_{lB_l})$$
$$= K_O 4^{-s} \prod_{l=1}^{s} [1+\exp(-\varepsilon_{l1})+ \exp(-\varepsilon_{l2})+\exp(-\varepsilon_{l3})] \quad (14a)$$
$$= K_O \prod_{l=1}^{s} q_l \quad (\lambda = 1),$$

where the sum is over all sequence combinations $\{B_l\}$ of length $s$. Thus, in the particular case when the selection parameter $\lambda = 1$, the statistics for a set of specific sites can be used directly to estimate the specific affinity of a random site. From equations (12b) and (14a), the average binding constant for a random site is:

$$K_R = K_O \exp(-\langle E\rangle_{seq} - I_{seq}); \quad \lambda = 1. \quad (14b)$$

This relationship between sequence statistics and specific affinity for a random site has been derived independently by Gary Stormo (University of Colorado; personal communication). As shown in the Appendix, equation (A33a), the relation (14b) holds also when base-pairs are not equiprobable in the genome; furthermore, it holds to first-order in $(\lambda-1)$ even when $\lambda$ is different from 1, see equation (A33b). Actually, $K_R$ is the average of the specific component of the binding constant for a random site. If the protein can also bind in a totally non-specific mode with binding constant $K_{ns}$, this constant should be added to equations (14a) and (14b) for an estimate of the overall non-specific binding constant (see von Hippel & Berg, 1986). If the protein binds to random DNA dominantly in this non-specific binding mode, $K_{ns} > K_R$ and $K_R$ may not be observable.

(iv) *Summary of the selection theory*

The results of the theory follow from the assumption that all base-pair sequences that provide the same specific affinity (or activity) have an equal probability of selection as recognition sites during evolution. To make the calculations more tractable, we have also added the assumption that individual base-pairs contribute independently to the affinity. As discussed further, below, together with the results of the sequence analyses, neither of these assumptions can be strictly true in general. However, the equiprobability assumption represents the simplest assumption that is consistent with what is known about neutral sequence drift and natural selection. The independence assumption is removed in the Appendix, so that the theory can account for the possibility that neighboring base-pairs contribute co-operatively to the binding affinity (or activity).

On the basis of these two assumptions the results follow from the calculations presented mostly in the Appendix. The statistical-mechanical analogy enables us to reduce the computations in the main text and to draw on various well-known results and concepts (e.g. Boltzmann factors, partition functions, etc.). It should therefore be stressed that this selection theory works in *analogy* with a statistical-mechanical ensemble. Specific sites are assumed to be selected to have affinity (or specific activity) in some useful range, while the possible states of a statistical-mechanical system are limited by the amount of energy that is available. This is why binding *free energies* of the DNA sequences serve as discrimination *energies* in the selection theory, in analogy to the energy levels for a statistical-mechanical system. Similarly (the negative of), the sequence information serves as the selection *entropy* describing the "degeneracy" (or sequence variability) of the sites, i.e. it provides an estimate of the number of different sequences that could possibly function as specific sites. The selection parameter $\lambda$ provides a coupling between the affinity requirement and the sequence variability. In effect, $\lambda$ is a coupling factor between the protein properties represented by the set of interaction free energies $\{\varepsilon_{lB}\}$ and the DNA properties in the form of the base-pair choices $\{f_{lB}\}$.

Sequence mutations that do not change the binding affinity very much are assumed neutral for selection so that all possible sequences with the required binding affinity are equiprobable. Thus, the sequence mutations are analogous to the thermal transitions in a statistical-mechanical system. A collection of specific sites, where each individual site has been selected to serve a somewhat different function, will *not* be strictly analogous to a statistical-mechanical ensemble where individual systems are interchangeable. This is why the functional distribution of sequence specificity, $g(E)$, in principle cannot be determined in analogy with a statistical-mechanical energy distribution. However, as we found above, the average sequence statistics are very insensitive to the actual form of the required functional specificity distribution so that the properties of individual sites can be analyzed properly.

In the Appendix the relations above have been calculated from first principles and have also been extended to account for the possibility that the base-pairs do not occur with equal probability in the genome. These relations cannot be expected to apply exactly like the corresponding statistical-mechanical relations because of the limitations in the number of possible realizations; the number of sequences conforming to some discrimination energy requirement may be of the order of $10^2$ to

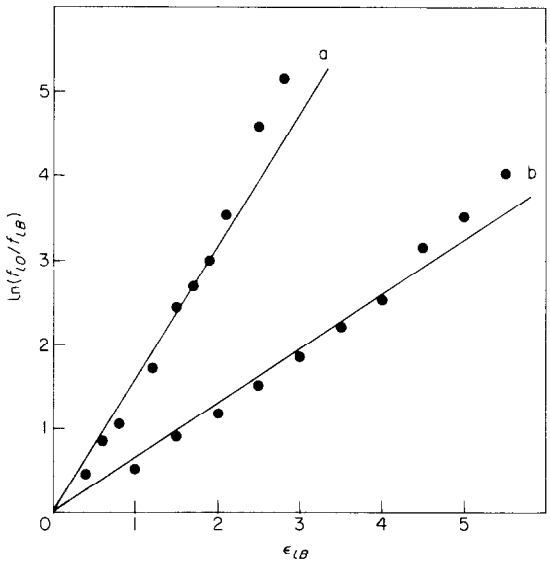


**Figure 1.** Relationship between base-pair frequencies $f_{lB}$ and discrimination energies $\varepsilon_{lB}$ in a randomly generated set of all sites with discrimination energy below a certain cut-off energy $E_c$. Data points are observed values and the lines are the expected relations using eqns (1) and (4) (or (A12) and (A13) from the Appendix). Slope a, results for $\{\varepsilon_{lB}\}_{l=1}^{10} = \{0{\cdot}4, 0{\cdot}6, 0{\cdot}8, 1{\cdot}2, 1{\cdot}5, 1{\cdot}7, 1{\cdot}9, 2{\cdot}1, 2{\cdot}5, 2{\cdot}8\}$ and $E_c = 3{\cdot}2$. At every position $l$, $\varepsilon_{l1} = \varepsilon_{l2} = \varepsilon_{l3}$ is assumed. A total of 706 different sequence combinations fall within the cut-off. Slope b, results for $\{\varepsilon_{lB}\}_{l=1}^{10} = \{1{\cdot}0, 1{\cdot}5, 2{\cdot}0, 2{\cdot}5, 3{\cdot}0, 3{\cdot}5, 4{\cdot}0, 4{\cdot}5, 5{\cdot}0, 5{\cdot}5\}$ and $E_c = 8{\cdot}7$. At every position $l$, $\varepsilon_{l1} = \varepsilon_{l2} = \varepsilon_{l3}$ is assumed. A total of 2029 different sequence combinations fall within the cut-off.

$10^6$ rather than $10^{20}$ as in a normal statistical-mechanical system.

We have also checked the relations on the computer by entering various sets of local discrimination energies $\{\varepsilon_{lB}\}$ and counting all possible binding sequences with an overall discrimination below some cut-off $E_c$. From the observed base-pair frequencies in the sample of all sites below this cut-off value, we can predict the local discrimination energies $\lambda\varepsilon_{lB}$ from equation (1) or equation (9) and compare them with the values $\varepsilon_{lB}$ that were assumed to generate the sequences. As seen from two representative examples in Figure 1, the basic result (eqn (1)) holds well except when $\varepsilon_{lB}$ approaches $E_c$. This limitation is expected since base-pairs with such large discriminations will not contribute in the sample of possible sequences. The straight lines in Figure 1 with a slope corresponding to $\lambda$ represent the predicted relation between the observed base-pair frequencies and the local discrimination energy. While the computer calculations quickly become too time-consuming when the site size and/or the number of different discrimination energies is large, we conclude from the results of Figure 1 that the theoretical relations above can be used for most situations and do, in fact, become more exact as the number of possible sequence combinations increases.

### (c) *Statistical small-sample errors*

In the preceding section we derived the expected relationship between discrimination energies and base-pair utilization frequencies in the set of all possible sites. The specific sites that have been experimentally identified and sequenced in any real case can be expected to form a very small subset of all sequences that could possibly have been used in nature. This will introduce a "small-sample effect", so that the observed base-pair utilization frequencies will not necessarily be identical to those expected. In the Appendix, equation (A40), we show that the best estimate of the true (in the set of all possible sites) base-pair frequency is:

$$f_{lB} = \frac{n_{lB}+1}{N+4}, \tag{15}$$

where $n_{lB}$ is the number of occurrences of base-pair $B$ at position $l$ in the sample of $N$ sites. This is often referred to as Laplace's Law of Succession. The assignment as in equation (15) makes obvious sense in the limit when no observations of the sequences have been made, i.e. when $n_{lB} = N = 0$, since under these conditions $f_{lB} = 1/4$ and all base-pair choices are equally probable at each position in the site. In other cases ($N \neq 0$), equation (15) provides the best assignment in the sense that it minimizes the uncertainty (or the expected variance) in the estimate. When used in place of the frequencies ($= n_{lB}/N$) actually observed in the sample, equation (15) smooths out the differences between the various base-pair choices somewhat and also affords a non-vanishing probability of occurrence to base-pairs that have not been observed in the sample. Although equation (15) gives the best estimate for the expected base-pair frequency $f_{lB}$, it does not necessarily give the best estimate for a function of $f_{lB}$ (like, e.g. $\lambda\varepsilon_{lB} = \ln(f_{l0}/f_{lB})$), as discussed in the Appendix. However, as a first approximation, our small-sample correction implies the usage of equation (15) in all expressions where $f_{lB}^{\text{obs}}$ is required.

With equations (9) and (15), the local discrimination energies should be estimated as:

$$\lambda\varepsilon_{lB} = \ln\left(\frac{n_{l0}+1}{n_{lB}+1}\right) \tag{16}$$

for each base-pair $B$ at position $l$.

In the Appendix the statistical errors in the base-pair frequency assignments introduced by the small sample have also been calculated. From equation (A41), the expected relative standard deviation $s_{lB}/f_{lB}$ in the base-pair frequency assignment is given by:

$$(s_{lB}/f_{lB})^2 = \frac{N+2-(n_{lB}+1)}{(n_{lB}+1)(N+5)} \approx \frac{1-f_{lB}}{Nf_{lB}} \quad (\text{for } N \gg 1). \tag{17}$$

Thus, the relative error is much smaller for base-pairs that occur frequently in the sample. For large

**Table 1**

*Small sample uncertainties and contributions from random fluctuations in base-pair choice at irrelevant positions*

| $N$ [a] | $s_{lB}/f_{lB}$ [b] | $\langle n_0 \rangle_R$ [c] | $n_0(5\%)$ [d] | $\langle \lambda\varepsilon \rangle_R$ [e] | $\langle I_l \rangle_R$ [f] |
|---|---|---|---|---|---|
| 10 | 0·45 | 4·2 | 7 | 0·30 | 0·078 |
| 20 | 0·35 | 7·4 | 11 | 0·27 | 0·055 |
| 30 | 0·29 | 10·4 | 14 | 0·25 | 0·042 |
| 40 | 0·26 | 13·4 | 17 | 0·24 | 0·033 |
| 50 | 0·23 | 16·1 | 21 | 0·22 | 0·027 |
| 60 | 0·21 | 19·1 | 24 | 0·20 | 0·022 |
| 70 | 0·20 | 21·9 | 27 | 0·19 | 0·019 |
| 80 | 0·19 | 24·5 | 29 | 0·18 | 0·018 |
| 90 | 0·18 | 27·2 | 33 | 0·17 | 0·013 |
| 100 | 0·17 | 30·2 | 36 | 0·17 | 0·014 |
| 112 | 0·16 | 33·5 | 40 | 0·16 | 0·013 |
| 120 | 0·15 | 35·8 | 42 | 0·16 | 0·012 |
| 200 | 0·12 | 57·5 | 65 | 0·13 | 0·0074 |

Columns c to f were calculated from 1000 randomly generated samples of size $N$.

[a] Sample size.

[b] Relative standard deviation in the assignment of base-pair frequency from eqn (17), calculated for an "average" base-pair observed $N/4$ times.

[c] Average number of occurrences of the most common base-pair at an irrelevant position.

[d] Number of occurrences of the most common base-pair, for which the probability is 5% or less, that a larger number will be observed at an irrelevant position. These numbers provide a quick estimate of the significance in base-pair variability.

[e] Average discrimination energy assigned to an irrelevant position.

[f] Average sequence information from eqn (A34) in a random assignment of base-pairs.

values of $N$, the measure in equation (17) agrees with the relative standard deviation in the frequency of occurrence of base-pair $B$ in the sample if it is known to occur with probability $f_{lB}$. In column b of Table 1 the relative error from equation (17) has been listed for various sample sizes $N$.

From equation (9) the expected standard deviation in $\lambda\varepsilon_{lB}$ would be approximately:

$$s_\varepsilon = [(s_{l0}/f_{l0})^2 + (s_{lB}/f_{lB})^2]^{\frac{1}{2}}. \quad (18)$$

From equations (11) and (16), the discrimination energy $E$ for a certain sequence $\{B_l\}_{l=1}^{s}$ would be estimated as:

$$\lambda E(\{B_l\}) = \sum_{l=1}^{s} \ln\left(\frac{n_{l0}+1}{n_{lB_l}+1}\right). \quad (19)$$

Since the variances are additive, the expected standard deviation in this estimate $\lambda E$ would be approximately:

$$s_E = \left\{\sum_{\substack{l=1 \\ B_l \neq 0}}^{s} [(s_{l0}/f_{l0})^2 + (s_{lB_l}/f_{lB_l})^2]\right\}^{\frac{1}{2}} \approx (6m/N)^{\frac{1}{2}}, \quad (20)$$

where $s_{lB}/f_{lB}$ is given by equation (17). Since the discrimination energy $E$ is defined relative to the consensus sequence, the sum in equation (20) is taken only over positions in the sequence at which base-pairs other than the consensus base-pair ($B = 0$) occur. In the approximate relation, $m$ is the number of non-consensus base-pairs in the sequence under consideration and this part of the expression has been evaluated using equation (17) with an average base-pair frequency $f_{l0} = f_{lB} = 1/4$. When two different sequences are compared, equation (20) gives the expected standard deviation of the difference in their discrimination levels ($\lambda E$) if the sum is taken instead only over the base-pairs that differ in the two sites.

Equation (20) accounts for the uncertainty in the discrimination energy relative to the consensus sequence. There is also an uncertainty as to whether the perceived consensus sequence really represents the cognate (best binding) sequence. As discussed in the Appendix, even irrelevant positions will be assigned positive (or possibly zero) discrimination energies through the use of equation (16). The inclusion of many irrelevant positions in the analysis will substantially increase the statistical uncertainty in the discrimination energies estimated from equation (19).

Equations (16) and (19) will serve as the basis for the sequence analyses below. The expected deviation given by equations (17) and (20) represent the statistical small-sample errors. If one observes deviations much larger than these, they are likely to represent errors in the physical assumptions, e.g. base-pairs at different positions that do not contribute independently to the binding affinity or certain highly conserved base-pairs in the sites that serve some purpose other than to contribute to the binding affinity. Such base-pairs will have a statistical weight in the sample that is not proportional to their effect on binding discrimination.

## 3. Sequence Analysis of Specific Sites

### (a) *Operator selection*

In principle, we can apply our theory directly *via* equation (16) to derive local discrimination energies ($\varepsilon_{lB}$) using published compendia of base-pair utilization frequencies; for example, repressor binding sites of various types on *E. coli* DNA. Base-pair frequencies in the binding sites for the *E. coli* *lexA*, *trpR*, *lacI* and *argR* gene products, as well as for the lambda *c*I and *cro* gene products have been assembled by Schneider *et al.* (1986). Unfortunately each of these sets of sites consists of less than a dozen DNA base-pair sequences. As a consequence of these small sample sizes the statistical errors are very large and predictions may be of limited value, especially if the binding constants of the relevant protein to the various sites differ rather little. Nevertheless, order-of-magnitude predictions can be made that are in reasonable accord with published binding data.

As an illustrative example we can consider the (non-co-operative) binding of the *cro* repressor of phage lambda to the two sets of three adjacent operator sites at the lambda $P_R$ and $P_L$ promoters.

**Table 2**

*Binding sites for* cro *and lambda* c*I repressors*

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Consensus | | T | T | A | T | C | A | C | C | G | G/C | C | G | G | T | G | A | T | A | A [a] |
| | | A | A | T | A | G | T | G | G | C | G/C | G | C | C | A | C | T | A | T | T |
| ($n_{1B}+1$) | A | 2 | 2 | 13 | 4 | 1 | 10 | 1 | 1 | 1 | 2 | 3 | 5 | 2 | 3 | 1 | 7 | 1 | 10 | 9 |
| | C | 3 | 3 | 1 | 4 | 13 | 2 | 12 | 8 | 2 | 6 | 10 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 2 |
| | G | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 10 | 6 | 2 | 8 | 12 | 2 | 13 | 4 | 1 | 3 | 3 [b] |
| | T | 9 | 10 | 1 | 7 | 1 | 3 | 2 | 5 | 3 | 2 | 1 | 1 | 1 | 10 | 1 | 4 | 13 | 2 | 2 |
| $O_{R3}$ | | C | T | A | T | C | A | C | C | G | C | A | A | G | G | G | A | T | A | A [c] |
| | | G | A | T | A | G | T | G | G | C | G | T | T | C | C | C | T | A | T | T |
| | | 3 | | | | | | | | | | 3 | 5 | | 2 | | | | | |
| | | 9 | | | | | | | | | | 10 | 8 | | 10 | | | | | 0·0125 [d] |
| $O_{R2}$ | | C | T | A | A | C | A | C | C | G | T | G | C | G | T | G | T | T | G | A [c] |
| | | G | A | T | T | G | T | G | G | C | A | C | G | C | A | C | A | A | C | T |
| | | 3 | | | 4 | | | | | | 2 | 2 | 2 | | | | 4 | | 3 | |
| | | 9 | | | 7 | | | | | | 6 | 10 | 8 | | | | 7 | | 10 | 5·4 $10^{-4}$ [d] |
| $O_{R1}$ | | T | T | A | C | C | T | C | T | G | G | C | G | G | T | G | A | T | A | A [c] |
| | | A | A | T | G | G | A | G | A | C | C | G | C | C | A | C | T | A | T | T |
| | | | | | 4 | | 3 | | 5 | | | | | | | | | | | |
| | | | | | 7 | | 10 | | 8 | | | | | | | | | | | 0·107 [d] |
| $O_{L1}$ | | A | T | A | C | C | A | C | T | G | G | C | G | G | T | G | A | T | A | C [c] |
| | | T | A | T | G | G | T | G | A | C | C | G | C | C | A | C | T | A | T | G |
| | | 2 | | | 4 | | | | 5 | | | | | | | | | | | 2 |
| | | 9 | | | 7 | | | | 8 | | | | | | | | | | | 9 | 0·0176 [d] |
| $O_{L2}$ | | T | T | A | T | C | T | C | T | G | G | C | G | G | T | G | T | T | G | A [c] |
| | | A | A | T | A | G | A | G | A | C | C | G | C | C | A | C | A | A | C | T |
| | | | | | | | 3 | | 5 | | | | | | | | 4 | | 3 | |
| | | | | | | | 10 | | 8 | | | | | | | | 7 | | 10 | 0·0321 [d] |
| $O_{L3}$ | | T | A | A | C | C | A | T | C | T | G | C | G | G | T | G | A | T | A | A [c] |
| | | A | T | T | G | G | T | A | G | A | C | G | C | C | A | C | T | A | T | T |
| | | | 2 | | 4 | | | 2 | | 3 | | | | | | | | | | |
| | | | 10 | | 7 | | | 12 | | 10 | | | | | | | | | | 5·7 $10^{-3}$ [d] |

[a] Symmetric consensus sequence formed by taking the most common base-pair at each position.

[b] Numbers of occurrences plus one ($n_{1B}+1$) for each base-pair at each position. Since the protein binds symmetrically, individual binding sites have been counted in both directions, thereby artificially increasing the sample from 6 to 12 (cf. Schneider *et al.*, 1986).

[c] The 6 binding sequences for *cro* and lambda *c*I repressor as listed, e.g., by Ohlendorf *et al.* (1982).

[d] The fractional number under a certain base-pair is $(n_{1B}+1)/(n_{10}+1) = \exp(-\lambda\varepsilon_{1B})$, which expresses the reduction in binding constant (taken to power $\lambda$) from a non-consensus base-pair in the sequence. The number on the right is the product of these reductions, which gives the total reduction in binding constant (taken to the power $\lambda$) for a particular sequence relative to the consensus sequence.

Taking these operator sites as being 19 base-pairs in length, we can use the base-pair utilization frequencies listed by Schneider *et al.* (1986), together with equation (19), to predict the relative binding constants (taken to the power $\lambda$) of these sites for *cro* repressor. The relative values of $\exp(-\lambda E)$ obtained for these sites, listed in the order $O_{R3}/O_{R2}/O_{R1}/O_{L1}/O_{L2}/O_{L3}$, are found to be: 1·0/0·044/8·6/1·4/2·6/0·46 (see Table 2). This set of relative values can be compared to the relative values of the measured binding constants (tabulated by Ohlendorf *et al.*, 1982), which are: 1·0/0·12/0·12/0·5/0·5/0·1. Obviously, no single value of $\lambda$ can be used that will make the predicted ratios agree with those observed. However, of the differences listed only the discrepancy factor of ~70 for the $O_{R1}$ site is really significant; the others fall within the expected statistical uncertainty of about a factor of 6 for a sample of this size (cf. eqn (20) and Table 1).

The discrepancy at $O_{R1}$ may have biological significance. The $O_{R1}$ operator is the strongest binding site (of the 6 *cro* operators under consideration) for the lambda *c*I repressor; thus, clearly its base-pair sequence has been selected to satisfy another strong constraint in addition to *cro* protein binding. The *c*I binding interaction involves significant contacts with the middle base-pairs of the operator sequences, while *cro* protein binding does not seem to involve these positions (see Ohlendorf *et al.*, 1982). If we recalculate our predicted ratios of binding constants, using only the 14 base-pairs (the central 17 base-pairs of the 19 base-pair sequence of each operator, omitting the central 3 base-pairs) that have been implicated in *cro* binding, we find for the expected ratios of $\exp(-\lambda E)$: 1·0/0·20/0·87/2·8/0·26/0·15. This brings all the calculated ratios of binding constants (assuming that $\lambda$ is close to unity) within or close to the expected standard deviation of about a factor of 6 from the ratios of the experimental values.

This result demonstrates that while predictions

may be of limited direct usefulness for sets of binding sites based on such small sample sizes, ratios of binding constants can be estimated at least to within an order of magnitude. Furthermore, the theoretical description is consistent with experiment, and large deviations from the predicted ratios can be used to infer the existence of other selection constraints that perturb base-pair utilization frequencies. In the following section we analyze *E. coli* promoter sequences to show that the theory can be used to make predictions of considerable utility for systems based on larger sample sizes.

### (b) *Promoter selection*

The initiation of transcripts by RNA polymerase at promoters, unlike repressor binding, is not an equilibrium selection process as described in the previous sections. However, it is a useful example for our selection theory because there are many more sequences available for promoters than for any other type of protein binding site on DNA. Hawley & McClure (1983) have compiled a list of 112 different promoter sequences from *E. coli*. We can extend the arguments for equilibrium selection to steady-state selection in the following way.

#### (i) *Promoter activity*

In a system with a collection of different (and non-interfering) promoters $P_i$ $(i = 1, 2, \ldots)$ the reaction scheme for chain initiation at each of them can be written as (cf. McClure, 1985):

$$P_i + R \underset{k_d^i}{\overset{k_a^i}{\rightleftarrows}} (P_iR)_{cl} \xrightarrow{k_2^i} (P_iR)_{op} \xrightarrow{k_3^i} P_i + R' \qquad (21)$$

where R denotes RNA polymerase, $(P_iR)_{cl}$ represents the initial (closed) complex of promoter and polymerase, and $(P_iR)_{op}$ denotes the "melted-in" (open) complex. The third step, $k_3^i$ is the rate with which the promoter is "cleared" by the elongating polymerase (R′), and thus made available to accept a new polymerase. Thus, the steady-state chain-initiation flux for each promoter of type $i$ in the system can be calculated as:

$$j_i = \frac{k_2^i K_B^i [R_f]}{1 + K_B^i [R_f](1 + k_2^i/k_3^i)}, \qquad (22)$$

where $[R_f]$ is the concentration of free polymerase and:

$$K_B^i = \frac{k_a^i}{k_2^i + k_d^i} \qquad (23)$$

corresponds to the inverse of the Michaelis–Menten constant. The ratio of initiation fluxes through different promoters is given by the ratio of their respective $k_2 K_B$ values if the denominator in equation (22) is close to unity, i.e. if the promoters are not saturated. There are strong indications that this may be the case *in vivo* (Bremer & Dalbow, 1975; Crooks *et al.*, 1983). Thus, the discrimination for promoter selection under steady-state conditions is determined by a ratio of $k_2 K_B$ values for the various promoter sequences in question, just as the equilibrium binding distribution is determined by ratios of binding constants. The discrimination free energies of our theory can then be replaced by a combination of binding and activation free energies.

#### (ii) *Promoter homology*

A correlation between *in vitro* values of $k_2 K_B$ and sequence has been demonstrated by Mulligan *et al.* (1984), who find a linear relation between $\log (k_2 K_B)$ and a "homology score" defined in terms of the deviation of each promoter from the consensus sequence. We are now in a position to apply physical theory to describe these correlations. The discrimination level $\lambda E$ defined in equation (19) will serve as a measure of the departure from homology. One basic difference of this approach from the *ad hoc* homology score defined by Mulligan *et al.* (1984) is that their score is derived by adding the observed base-pair frequencies $n_{lB_l}$, while the measure $\lambda E$ from equation (19) adds $\ln (n_{lB_l} + 1)$ for every base-pair $B_l$ in a sequence $\{B_l\}_{l=1}^{s}$.

To apply our theory, we must first include the contributions from the variable-length spacer region between the two important sequence regions around positions $-10$ and $-35$. In agreement with previous assumptions we assume that the spacer contributes independently to the binding interactions. Then the variation in spacer length can be shown to contribute an additive term to the overall discrimination energy (see eqn (A21) of the Appendix). As a consequence, equation (19) becomes:

$$\lambda E(\{B_l\}, L) = \sum_{l=1}^{s} \ln\left(\frac{n_{l0}+1}{n_{lB_l}+1}\right) + \ln\left[\frac{n(L_{opt})+1}{n(L)+1}\right], \qquad (24)$$

where $L_{opt}$ ($= 17$ for the promoters) is the optimal spacer length and $L$ is the actual spacer length for the sequence in question. $n(L)$ is the observed number of occurrences for spacer length $L$ in the sample of sequenced promoters. Using equation (24) requires that every given specific sequence be aligned with the consensus sequence in only one way; otherwise the number of occurrences $n(L)$ of a certain spacer length is not uniquely defined. Thus, both the $-10$ and the $-35$ regions must be so well-defined that alternative alignments (assuming different spacer lengths) are not possible. This is true for most of the promoter sequences listed, but certainly not for random sequences. In principle it would be possible to relax this assumption and enter different alignments with different weights.

#### (iii) *Activity–homology correlations*

Using the compilation of base-pair frequencies obtained by Hawley & McClure (1983), the 30-base-pair site size, and the list of *in vitro* $k_2 K_B$ values for 31 promoters presented by Mulligan *et al.* (1984), we find the correlation plotted in Figure 2 between $\ln (k_2 K_B)$ and $\lambda E$ defined by equation (24). A least-squares line can be fitted fairly well through the data points with a correlation coefficient $r = 0{\cdot}84$.

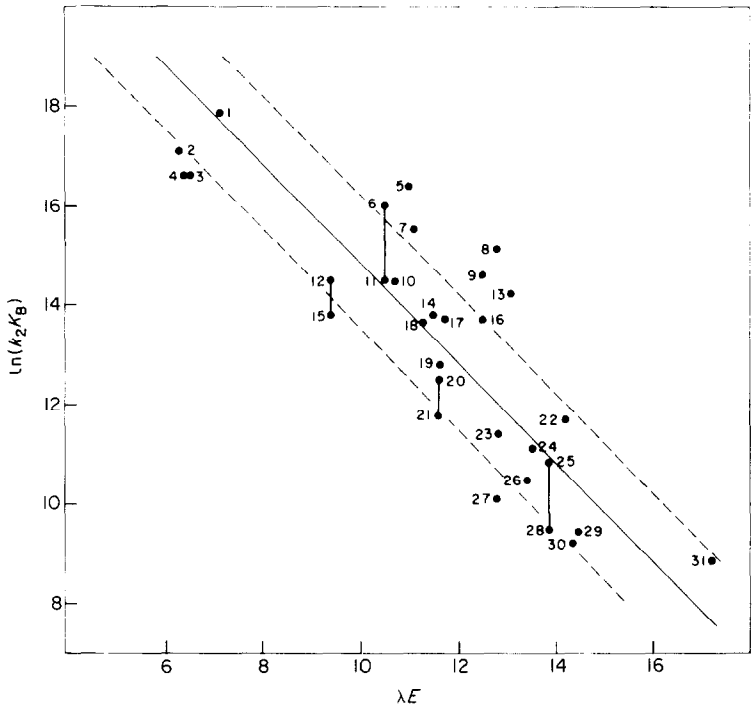


**Figure 2.** Observed correlation between calculated discrimination level $\lambda E$ and *in vitro* activity $k_2K_B$ for various promoters. Since there are uncertainties in both co-ordinates ($\lambda E$ and $\ln(k_2K_B)$) in this Figure, the least-squares line ($\ln(k_2K_B) = -1{\cdot}00\lambda E + 24{\cdot}85$) has been determined by minimizing the average perpendicular distance from the data points. The numbers for the various promoters and references to the original literature are as given by Mulligan *et al.* (1984). The broken lines represent a $\pm 1$ standard deviation, both as observed for the 31 data points and as expected from the predicted uncertainty ($\sim \pm 1$ unit) in $\lambda E$ and the experimental uncertainty in $\ln(k_2K_B)$.

This is only a marginally better fit than was obtained with the homology score used by Mulligan *et al.* (1984), which gave $r = 0{\cdot}83$. From our theory we expect that $k_2K_B$ will vary proportionally to $\exp(-E)$. Thus, the slope of the line in Figure 2 gives the parameter $\lambda \approx 1{\cdot}0$, although there is a large uncertainty in the quantitative estimate of the slope interpreted as the physical parameter $\lambda$. In Table 3 we have listed the observed base-pair frequency data from Hawley & McClure (1983) and some of the various quantities that are relevant for the statistical-mechanical analysis. In keeping with the basic selection assumption of the theory we have excluded the six promoter sequences listed by Hawley & McClure (1983) that were created by fusion or mutation, and have considered in the basic set only the 106 naturally occurring sequences. This exclusion has a very small effect on the numerical results.

The large statistical uncertainties in the estimated discrimination energy discussed in section 2(c), above (cf. Table 1 and eqn (20)) lead to a standard deviation in the estimated numerical value for $\lambda E$ of about $\pm 1$ unit; the uncertainty is somewhat smaller for sites with good homology. There is also a large experimental uncertainty in the promoter strengths as given by $k_2K_B$. In Figure 2, data points connected by a continuous vertical line correspond to measurements of $k_2K_B$ for the same promoter carried out in different laboratories (see Mulligan *et al.*, 1984). Thus, agreement with the theory can be expected only on average. In fact, the deviations from the least-squares line observed in Figure 2 are of the magnitude expected from the uncertainties in the two co-ordinates.

Extending the straight line in Figure 2 to maximum homology (discrimination $E = 0$) would give $(k_2K_B)_{max} = 10^{11}\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$. However, there is no reason to expect the linear relationship to extend that far. From column d of Table 3 it can be estimated that between four and nine positions in the site size may be irrelvant for specificity since the base variations at these positions could well be caused by random fluctuations. Each of these irrelevant positions would contribute, on average, 0·16 (see Appendix, section (c), and column e of Table 1) to the estimate for $\lambda E$. Thus, $\lambda E$ could probably not be smaller than about 1.

More important, however, is the fact that $k_2K_B$ is a combination of kinetic factors. From equation (23) we have $k_2K_B = k_ak_2/(k_2+k_d)$. Most of the promoters in Figure 2 probably work in the limit where $k_d \gg k_2$, so that the observed correlation between sequence homology and activity $k_2K_B$ actually pertains to $k_2K_B \approx k_ak_2/k_d$. Thus, if $k_a$ is fairly insensitive to sequence (e.g. diffusion-limited) the observed correlation is between sequence homology and the ratio $k_2/k_d$. When this ratio becomes large, promoter activity will become association limited, $k_2K_B \approx k_a$ from equation (23). Then the straight line in Figure 2 should level off at

**Table 3**
*Statistics from the promoters*

| [a] | [b] | A | C | G | T[c] | $I_l$[d] | $\lambda\langle\varepsilon_l\rangle$[e] | [f] | $n_2^{obs}:\bar{n}_2$[g] | [h] |
|---|---|---|---|---|---|---|---|---|---|---|
| −45 | A | 52 | 15 | 19 | 23 | 0·127 | 0·519 | AA | 23 : 18·9 | 1·0 |
| | A | 40 | 20 | 14 | 35 | 0·079 | 0·305 | AA | 19 : 15·6 | 0·9 |
| | A | 43 | 12 | 20 | 34 | 0·102 | 0·354 | AA | 20 : 14·4 | 1·6 |
| | T | 37 | 21 | 10 | 41 | 0·115 | 0·293 | AA | 19 : 13·0 | 1·8 |
| | A | 39 | 23 | 20 | 27 | 0·035 | 0·325 | AA | 14 : 13·4 | 0·2 |
| −40 | A | 38 | 17 | 24 | 30 | 0·041 | 0·292 | TT | 14 : 10·5 | 1·1 |
| | T | 19 | 26 | 25 | 39 | 0·034 | 0·324 | TT | 16 : 15·9 | 0·0 |
| | T | 26 | 11 | 27 | 45 | 0·102 | 0·400 | TA | 15 : 10·1 | 1·6 |
| | C | 25 | 41 | 29 | 15 | 0·059 | 0·341 | CT | 31 : 32·5 | −0·3 |
| | T | 3 | 8 | 12 | 87 | 0·670 | 0·482 | TT | 75 : 73·0 | 0·4 |
| −35 | T | 6 | 7 | 6 | 91 | 0·737 | 0·460 | TG | 71 : 72·2 | −0·3 |
| | G | 3 | 11 | 86 | 10 | 0·647 | 0·493 | GA | 56 : 55·3 | 0·1 |
| | A | 70 | 18 | 3 | 19 | 0·401 | 0·533 | AC | 44 : 37·8 | 1·3 |
| | C | 25 | 59 | 11 | 15 | 0·213 | 0·550 | CA | 32 : 26·3 | 1·3 |
| | A | 49 | 9 | 17 | 35 | 0·168 | 0·409 | TT | 21 : 13·8 | 2·1 |
| −30 | T | 26 | 25 | 15 | 44 | 0·070 | 0·400 | | | |
| | | | | (spacer) | | | 0·694 | | | |
| | T | 22 | 26 | 14 | 48 | 0·099 | 0·458 | TA | 16 : 12·4 | 1·1 |
| | T | 29 | 19 | 29 | 33 | 0·019 | 0·163 | AT | 17 : 10·6 | 2·1 |
| | T | 17 | 25 | 27 | 41 | 0·048 | 0·351 | TG | 28 : 15·5 | 3·4 |
| −15 | G | 23 | 20 | 42 | 25 | 0·045 | 0·379 | GT | 15 : 11·2 | 1·2 |
| | G | 18 | 27 | 35 | 30 | 0·027 | 0·215 | GT | 29 : 26·6 | 0·5 |
| | T | 3 | 11 | 12 | 84 | 0·610 | 0·506 | TA | 77 : 78·3 | −0·3 |
| | A | 101 | 3 | 2 | 4 | 1·016 | 0·285 | AT | 43 : 44·3 | −0·3 |
| | T | 28 | 16 | 18 | 48 | 0·100 | 0·457 | TA | 29 : 27·5 | 0·3 |
| −10 | A | 63 | 15 | 18 | 14 | 0·237 | 0·592 | AA | 33 : 31·6 | 0·3 |
| | A | 55 | 22 | 14 | 19 | 0·152 | 0·541 | AT | 54 : 52·0 | 0·4 |
| | T | 2 | 4 | 1 | 103 | 1·089 | 0·232 | TA | 32 : 31·6 | 0·1 |
| | A | 34 | 14 | 33 | 29 | 0·048 | 0·164 | GC | 19 : 10·9 | 2·6 |
| | C | 22 | 37 | 22 | 29 | 0·025 | 0·272 | CG | 15 : 9·5 | 1·9 |
| −5 | A | 30 | 30 | 29 | 21 | 0·010 | 0·077 | | | |

[a] Position number in the promoter sites as labeled by Hawley & McClure (1983).
[b] Consensus sequence.
[c] Base-pair utilization $(n_{lB}+1)$ for $B$ = A, C, G, T, at position $l$.
[d] Sequence information at position $l$ from eqn (A35). From eqn (A36) $\exp(-NI_l)$ gives a measure for the probability of random occurrence of the observed base-pair utilization at this position. The sum of the entries in this column gives $I_{seq} = 7{\cdot}1$.
[e] $\langle\varepsilon_l\rangle$ is the average contribution to the discrimination energy at this position. The sum of the entries in this column gives $\lambda\langle E\rangle_{seq} = 12{\cdot}0$.
[f] Most frequent doublet at this position and the following.
[g] $n_2^{obs}$ is the number of occurrences of the most frequent doublet and $\bar{n}_2$ is the expected number based on the singlet frequencies of the respective base-pairs.
[h] $(n_2^{obs}-\bar{n}_2)/[\bar{n}_2(1-n_2/\bar{N})]^{\frac{1}{2}}$ is the deviation in the observed doublet frequency divided by the expected standard deviation. This gives a measure of the significance of the observed doublet correlation.

$k_2k_B \approx k_a$ when the discrimination energy decreases below that which corresponds to $k_d/k_2 \sim 1$. In fact, the "best" promoters in Figure 2 may already be approaching this limit. This may also be the reason why a very efficient synthetic promoter with close to maximal homology exhibits almost no change in its *in vitro* activity when one of the strongly conserved base-pairs in the −35 region is substituted (Rossi *et al.*, 1983). This substitution would increase the discrimination level $\lambda E$ from 1·1 to 2·8. These small values would keep both sites well within the region conjectured to correspond to an association-limited promoter activity (cf. Fig. 2).

Similarly, when the discrimination energy becomes very large (and $k_2K_B$ becomes very small) other kinds of interactions (e.g. purely non-specific electrostatic) may become dominant so that the straight line cannot be extended too far in this direction either. Also, since the slope is rather uncertain even within the range of the available data, predictions too far outside the observed range would not be very reliable even if the linearity observed does hold over the entire range.

(iv) *Promoter classification*

If the promoters are classified according to increasing values of $\lambda E$ (see Fig. 3), their order will differ appreciably (in detail) from the list given by Mulligan *et al.* (1984) for decreasing values of their homology score. This is not surprising, since the homology score is defined quite differently from $\lambda E$ given by equation (24). However, the gross features of these classifications are very similar within the expected statistical errors, so that a promoter with a large homology score according to Mulligan *et al.* (1984) will have a small discrimination level $\lambda E$ in

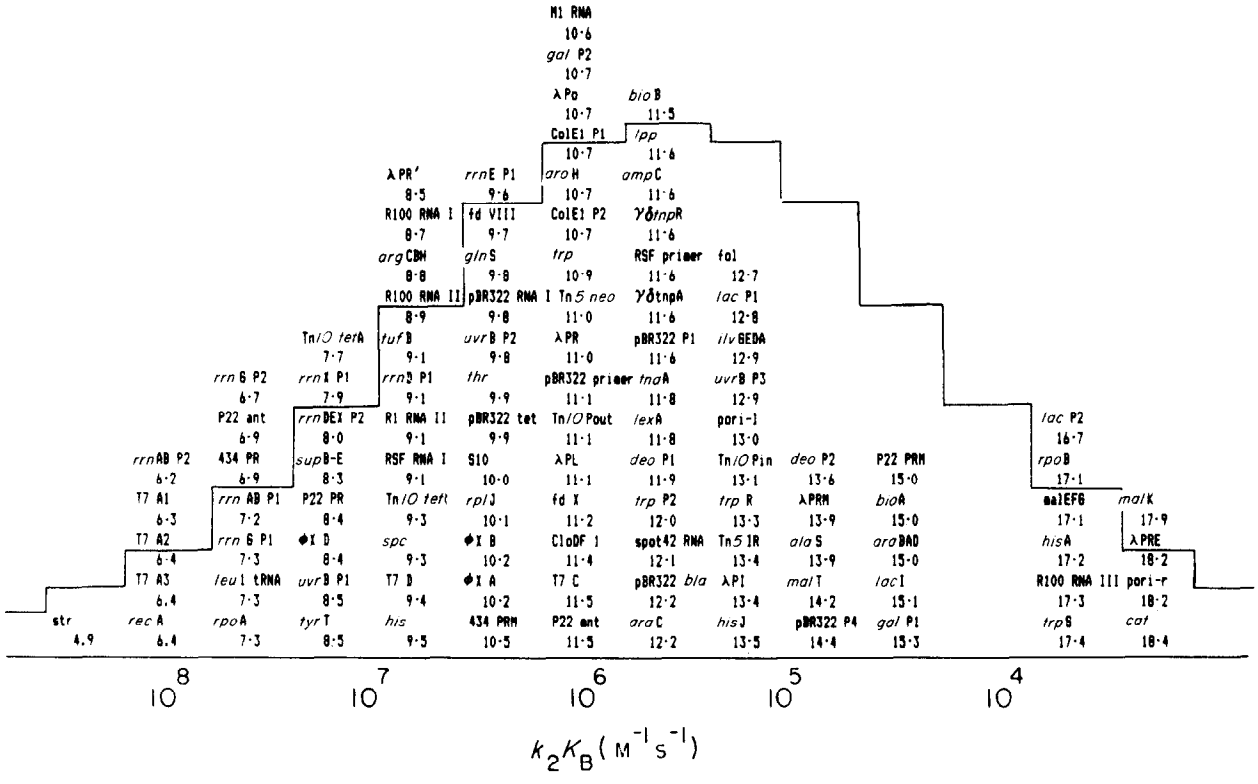


**Figure 3.** The promoters in the sample, rank-ordered in accord with their discrimination level $\lambda E$, which is given below each promoter name. To see the form of the distribution of promoters along the discrimination level co-ordinate, the promoters have been put into groups for which $\lambda E$ differs by 1 unit. Promoters with half-interger values of $\lambda E$ at the separation between groups have been put arbitrarily in the group to which they would belong when second decimals are included in the value of $\lambda E$. For comparison, the distribution that would be expected if the promoters conform to a canonical ensemble has also been drawn (continuous line), and is approximated as a Gaussian with variance $\lambda^2\sigma_E^2 = 10{\cdot}4$ as determined from eqn (A32). Since $\lambda \approx 1$ for the promoters, in accordance with eqn (A20) this curve approximately describes also the expected distribution of random activity in the genome at pseudosites with the discrimination levels indicated. Numbers on the bottom are the predicted *in vitro* activities, if the correlation observed in Fig. 2 holds.

our listing. In fact, a linear relationship between the two measures holds very well (with correlation coefficient 0·98) over the entire range of the promoter sample. However, this linear relationship between the two measures relies on an approximate proportionality between

$$\sum_l \ln\,[(n_{l0}+1)/(n_{lB_l}+1)] \quad \text{and} \quad \sum_l (n_{l0}-n_{lB_l}).$$

This is certainly not true in general, and may not hold as well for arbitrary sequences or different samples. Thus, since equation (24) is based on physical theory, we propose that the discrimination level $\lambda E$ provides a better and more general measure for the departure from sequence homology as it pertains to binding and activity.

The small-sample uncertainties in the correlations between $\lambda E$ and the $k_2K_B$ values makes a detailed classification of predicted promoter strength according to sequence homology impossible, although a gross classification should work. This statistical uncertainty also implies that no strict cut-off in the value of $\lambda E$ (or any other homology score based solely on sequence data) can exist that sharply separates promoter from non-promoter sites.

### (c) *Conclusions from the sequence analysis*

The theory requires that the functional DNA sites have been selected according to some constraint (e.g. the binding affinity or activity must be in some useful range), and that this constraint operates for all sites. Obviously, problems of interpretation will occur when many sites have been selected to satisfy other criteria simultaneously, for example, promoter sites that also bind effector molecules, etc.

#### (i) *Base-pair independence*

As developed and applied above, the theory also requires that base-pair substitutions act independently, so that their contribution to the interaction free energy is additive. There are physical reasons to expect that this is not generally true (cf. von Hippel & Berg, 1986). First, it is likely that recognition is affected by the secondary effects that base sequences have on the local DNA structure and flexibility; since these DNA properties are determined primarily by interactions between neighboring base-pairs (Dickerson, 1983), if such effects dominate the specificity they would be expected to lead to strong correlations in the base-

pair usage at neighboring positions. Furthermore, even if specificity were based solely on the complementarity of the hydrogen-bond-forming groups of the functional sites of the protein and DNA, one might expect that the loss of contiguous specific contacts will contribute differently from the loss of non-contiguous ones (cf. Mossing & Record, 1985). However, as we shall discuss below, the statistics of base-pair usage in the promoter sample do not indicate that such co-operative effects have a dominant influence on polymerase recognition.

In principle, the theory can be extended to include correlations and co-operativity between different base-pairs (see the Appendix). However, because of the much larger range of possibilities for doublets, triplets, etc., more sequence data than is available at present are required to calculate such correlations reliably. To estimate the importance of such correlations in the promoter sample, we have counted the occurrences of all doublets of neighboring base-pairs and compared them to the doublet frequencies expected from the single base-pair frequencies at the two positions in question (cf. columns f and g of Table 3). The strongest correlation is at three and two base-pairs upstream from the TATA box (i.e. positions −16 and −15 as numbered in Table 3), where T and G, respectively, are weakly preferred. Here, the doublet T-G occurs in 28 cases while only 15 to 16 would have been expected from the singlet occurrences of base T and base G at the respective positions. This is a highly significant deviation from individual base-pair independence, in that it would have occurred at random with a probability of only $6\times10^{-4}$. This result suggests that when either T or G is substituted, the choice of base-pair at the neighboring position becomes irrelevant.

A nearest-neighbor correlation like this can come about in several ways. It could reflect a physical interaction, i.e. that the polymerase can only make good contact with the DNA site when both T and G are present. Alternatively, it could reflect a subset of promoters requiring T-G as a signal, either for the binding of the polymerase or for the binding of some effector. There is a similar (though weaker) preference for an A-T doublet just upstream at positions −17 and −16. However, there is no significant preference for the simultaneous presence of these doublets, i.e. for the triplet ATG at these positions. Since neither of these doublets shows strong correlations with neighboring base-pairs on either side, it is unlikely that these positions serve as a signal related to the binding of an effector molecule other than the polymerase.

Partially overlapping these doublets there is also a preference for CTC at positions −18 to −16, which shows up as strong doublet correlations for C-T and T-C at their respective positions. Out of a total of nine occurrences of this triplet, five are found among the 18 rRNA and tRNA promoters. Thus, this triplet could serve as part of a signal defining a certain class of promoters. Alternatively, it could reflect a close relationship, i.e. perhaps some of these promoters have only recently evolved from the same common ancestor.

The second strongest doublet correlation occurs directly downstream from the conserved −10 region. Here (at position −7 and −6) the doublet G-C occurs 19 times, while only 11 would be expected from the respective singlet frequencies (see Table 3). Of these 19 occurrences, however, 14 derive from the 18 rRNA and tRNA promoters in the sample. This doublet is the first part of the discriminator region of sequence GCGC that is required for stringent control of stable RNA synthesis (Lamond, 1985). It is interesting to note that this correlation occurs in a region where the base-pair choice on the singlet level seems random, thus strengthening the suggestion that this signal is not directly related to polymerase activity.

There are also two weaker correlations surrounding the conserved −35 region. At positions −42−−41 and −31−−30 T-A and A-T, respectively, are unfavored while T-T and A-A are slightly favored in both cases (see Table 3).

Similarly, we have looked at the correlations between next-nearest and next-next-nearest-neighbors and find only a few, all of which are connected with the doublets discussed above. There is a relatively strong preference for A and G at positions −9 and −7 with 24 occurrences rather than the 16 expected from the singlet data. This correlation disappears when the 18 rRNA and tRNA promoters are excluded from the sample. This does not necessarily imply that the A should be considered as part of the signal in the discriminator region; possibly, it simply reflects the fact that most stable RNA promoters that carry the discriminator signal also are strong promoters, thus requiring the consensus A at position −9.

To gauge the importance of the doublet correlations observed, we have generated sets of "random promoter sequences" on a computer, where at every position a base-pair has been assigned in proportion to its frequency of occurrence in the promoter sample without regard to base-pair assignments at other positions. This procedure yields sequences where the single-base-pair occurrences agree approximately with those in the promoter sample, but where doublet correlations are due only to random small-number fluctuations. This provides a numerical "base-line" against which to assess the significance of the correlations found in the "real" promoter sequences.

We find that of the 28 doublet positions included in our study of the real promoters, only the six positions discussed above show significant correlations above the random variation expected. To quantify further the statistical significance of the doublet correlations, we have also calculated the "doublet information content" ($I_2$) as defined by equation (A38b) in the Appendix; for the real promoters it is $I_2 = 1{\cdot}6$ as compared to about 1·1 or 1·2 for the randomly generated ones. In contrast, the primary sequence information that measures the importance of the individual base-pair choices

in the promoter sample is around $I_{seq} = 7$ (see Table 3), while for a similar sample of totally random sequence it is $I_{seq} = 0{\cdot}4$ (see Table 1).

Thus, a few significant nearest-neighbor correlations exist in the promoter sequences, and some of these no doubt derive from co-operativity in the interactions between the polymerase and the individual base-pairs. Others seem to be part of signals that are not directly related to polymerase activity and therefore should not be ascribed to co-operativity. At most positions, however, the observed doublet correlations are not distinguishable above the small-number fluctuations, and base-pair occurrences seem, indeed, to be largely independent. This does not prove that the contributions from different base-pair positions to the interaction free energies are additive, although it is an indication that additivity is dominant and a reasonable first approximation. Apart from the few cases discussed above, the statistics of doublet occurrences in the promoter sample do not suggest any major revision of this assumption. To really prove independence and additivity would, of course, require experimental verification by systematic base-pair substitution; the statistics can suggest where deviations are most likely to occur.

It should be noted that the discrimination energies calculated from the single base-pair occurrences already include some average of possible co-operative effects, e.g. a strong co-operativity between two neighboring base-pairs that leads to a selection for them as a doublet will also increase their singlet frequencies, even if they do not contribute to recognition individually. Thus, the effects of nearest-neighbor co-operativity will be to modulate the assigned discrimination energy for every possible base-pair, depending on its nearest neighbors in the site as given by equation (A38a) in the Appendix.

Applied to the promoter sample, this modulation has a relatively small effect on the assignment of $\lambda E$ for most promoters. The statistical correlation with *in vitro* activity becomes somewhat better than that depicted in Figure 2, with one glaring exception: the L305 mutations (no. 9 and no. 24 in Fig. 2) of the *lac*UV5 and *lac*P$^s$ promoters are pushed far away from the least-squares line, to $\lambda E \approx 16$ and $\lambda E \approx 17$, respectively. This mutant has a single base-pair deletion just upstream from the $-35$ region that presumably realigns a number of base-pairs in the relatively unimportant region further upstream. The discrepancy for the L305 mutant when doublet correlations are included may simply reflect the fortuitous addition of contributions from doublets that are individually statistically insignificant.

The introduction of doublet correlations into the discrimination calculations increases the statistical uncertainties, since it requires the addition of a large number of imprecise data points. The doublet correction will be useful mostly when the base-pair correlations are so large that their presence introduces systematic errors that are larger than the statistical uncertainties in the correction terms. Thus, it will be more useful to apply the doublet corrections only to cases where the statistics suggest that doublets are important. In the case of promoters, such an application has only a minor influence on the estimated discrimination energies.

Co-operativity and correlations can also be included in a systematic and useful way for samples that are not prohibitively large by using smaller "alphabets" to reduce the number of possible combinations. For example, if DNA structure is determined mostly by purine-pyrimidine (rather than individual base-pair) choices (Dickerson, 1983) one need only consider four possible doublets and 16 triplets, rather than the 16 and 64, respectively, that apply to the full DNA alphabet when one utilizes all four base-pairs.

(ii) *Functional distribution*

The selection of samples may be biased in several ways. For example, it may be easier experimentally to identify strong sequences among all those that are used in the genome. Nature may also be biased in its choice of real sites among the potentially useful ones. In principle we cannot, in our analysis, distinguish between such biases. However, from the results presented in the second section it can be expected that such biases will primarily influence the parameter $\lambda$, and will leave the basic relation (eqn (16)) between the discrimination factors and base-pair utilization frequencies otherwise essentially unchanged. Thus, the correlation between sequence and discrimination should be largely invariant, although $\lambda$ cannot be calculated *a priori*.

From the sequence analysis of the promoters it is also possible to determine their distribution $g(E)$ along the discrimination-level co-ordinate $\lambda E$. The promoter list as depicted in Figure 3 can be viewed as a bar-graph representation of this distribution; it can be regarded as a distribution over the primary sequence specificity. The numbers on the bottom are the predicted *in vitro* activities that apply if the correlation found in Figure 2 holds. If the concentration of free polymerase is $3 \times 10^{-8}$ M, as it is suggested to be *in vivo* (McClure, 1985), the distribution spans initiation frequencies from about $1\ s^{-1}$ to $1\ h^{-1}$, which seems a reasonable range. It should be stressed, however, that the functional activities *in vivo* will be influenced by many other factors (e.g. supercoiling, activator proteins, etc.) that could appreciably change the overall form of the activity distribution.

For comparison, in Figure 3 we have also plotted the canonical distribution discussed in the Appendix and approximated as a Gaussian with mean $\lambda\langle E\rangle_{seq}$ and variance $\lambda^2\sigma_E^2$ given by equations (8) and (A32), respectively. While this construction forces the means of the observed distribution and the canonical one to agree *via* equation (8), it is interesting to note that the *widths* of the two distributions also agree quite well. This may well be coincidental, but could also reflect some evolutionary advantage in selecting specific sequences

with a canonical bias. As discussed in the Appendix, the canonical distribution confers maximal sequence variability within the constraint set by the maintenance of a given average discrimination energy $\langle E \rangle_{seq}$. Preliminary calculations (O. G. Berg, unpublished results) on 117 ribosome initiation sites (Gold *et al.*, 1981) show that these sites similarly conform to a canonical distribution over the discrimination level $\lambda E$. This may be an indication that sequence variability (or sequence diversity) is of primary importance in the evolutionary selection of recognition sequences.

### (d) *Sequence information and overspecification of binding sites*

Schneider *et al.* (1986) have used information theory to calculate the sequence information (defined from the observed base-pair frequencies *via* equations (12) and (A34)) for the binding sites of various recognizer proteins. Their analysis demonstrates the usefulness of sequence information in assessing the relative importance of various positions in the site; notably, it can be used to delineate those positions within the binding sequence that are really relevant *via* equation (A35). However, as shown above and in the Appendix, it is not the sequence information *per se*, but the ratio of base-pair frequencies that is directly related to the free energy of binding. While the sequence information is a measure for the whole set of sites, the statistical-mechanical sequence analysis can also provide a quantitative measure of specificity for individual sites.

#### (i) *Operator sites*

Schneider *et al.* (1986) found, for all the sets of specific sites investigated by them, that the sequence information is approximately equal to the negative logarithm of the probability that a site chosen at random in the genome is a specific site. Since the information content is essentially the negative logarithm of the probability of random occurrence of a potential site of average binding strength or greater (cf. eqn (13), one would conclude that the number of specific sites in the genome is approximately equal to the expected number in a random genome of the same size. However, the reduction factor (the denominator in eqn (13)) could reduce the estimate of randomly occurring binding sites by an order of magnitude or more; thus sequence information alone does not provide a reliable estimate of this expected frequency.

It has been argued (von Hippel, 1979) that specific sites in the genome should be specified in a way that makes the random occurrence of competitive binding sites ("pseudosites") unlikely. Without such "overspecification", the recognizer protein would be "soaked up" by binding to a large number of such pseudosites. In a recent paper (von Hippel & Berg, 1986) we showed quantitatively how binding selection can be balanced by sequence length (site size), discrimination factors and protein concentration. The amount of overspecification required is determined primarily by the number of protein molecules the system can afford to lose by non-productive binding at pseudosites. We can now relate this number quantitatively to the sequence information as follows. From equation (3) of von Hippel & Berg (1986) we can express the number of proteins bound at pseudosites as:

$$m_s = 2N_T F_A \sum_i p_s(E_i) \frac{x}{x + \exp(E_i - E_s)}, \quad (25)$$

where the sum is taken over all classes ($i$) of sites in the genome. $p_s(E_i)$ is the probability of random occurrence of a pseudosite with discrimination $E_i$, and $2N_T F_A$ is the total number of available binding sites in the genome. $N_T$ is the size of the genome in base-pairs and $F_A$ is a reduction factor that accounts for the fact that only a fraction of the genome may be available for binding; the rest may be covered by other proteins or structurally inaccessible for other reasons. The saturation level $x$ is defined from the fraction saturation $\theta_s$ of the specific site (with discrimination $E_s$) as $x \equiv \theta_s/(1-\theta_s)$. Since $p_s(E_i)$ increases rapidly with increasing discrimination $E_i$, the sum in equation (25) is dominated by terms for which $\exp(E_i - E_s) > x$. Then one finds:

$$m_s \approx 2N_T F_A x \exp(E_s) \times \sum_i p_s(E_i) \exp(-E_i) = 2N_T F_A x \exp(-I_{seq}). \quad (26)$$

The sum in equation (26) is the same as was calculated in equation (14a) and (14b) and the result holds if $\lambda \approx 1$ and if $x$ is determined by the fractional saturation of a specific site with average discrimination $E_s = \langle E \rangle_{seq}$. When the saturation effects of the pseudosites cannot be neglected, a correction factor ($<1$) should be included in equation (26). However, we find this to make a very small difference in all examples where we have summed equation (25) exactly; the non-specific competition from pseudosites is expected to be dominated totally by the large number of weak (unsaturated) sites rather than by a few strong ones. Thus, the observed sequence information in a set of binding sites can be related directly to the expected number of protein molecules wasted by non-productive binding at pseudosites. This number is modulated primarily by the saturation level $x$ required at an average specific site. Therefore, the relationship between sequence information and the number of specific sites in the genome is likely to be a complicated function that also involves the details of the regulatory requirements of the system. This will be discussed in more detail in a subsequent paper (O. G. Berg & P. H. von Hippel, unpublished results) in which we analyze the sequence specificity of the DNA binding sites for the cyclic AMP receptor protein (de Crombrugghe *et al.*, 1984).

In this connection it is also interesting to note, for the repressor binding sites studied by Schneider

*et al.* (1986), that the ones with the smallest sequence information are, at least in part, co-operative sites (e.g. sites for arginine repressor and lambda *c*I repressor) where binding to two neighboring sites is favored, while the ones with the largest sequence information are mostly independent sites (e.g. sites for tryptophan repressor and *lexA* protein). Obviously, when regulatory proteins bind co-operatively to two neighboring sites, less specificity is required for each individual binding site to achieve a required binding level. Thus sequence information can indeed be a useful measure for specificity, although one should be aware that it deals only with an *entropic* aspect of the specificity of a whole set of sites and cannot describe the specificity of individual sites.

(ii) *Promoter sites*

We can also estimate the number $N_s$ of randomly occurring promoter sites (pseudopromoters) in the genome using the promoter sample data:

$$N_s(\langle E\rangle_{seq}) = 2N_T P_s(\langle E\rangle_{seq}). \quad (27)$$

Using equation (13) for the probability $P_s(\langle E\rangle_{seq})$ of random occurrence of a site with discrimination less than $\langle E\rangle$ and data from Table 2 ($I_{seq} = 7{\cdot}1$, $\lambda\langle E\rangle_{seq} = 12{\cdot}0$, $\lambda\bar{\varepsilon} = 1{\cdot}2$, site size $s = 30$, and a genome size of $N_T = 10^7$), this gives $N_s(\langle E\rangle_{seq}) = 1000$. Accounting for various spacer lengths with an extra factor $1/f(L_{opt})$ from equation (A22) gives $N_s(\langle E\rangle_{seq}) = 2000$ for the number of pseudopromoters with discrimination less than the average in the sample of real promoters. The number of pseudopromoters with $\lambda E < 15$, which is where most of the real promoters fall (cf. Fig. 3), would be somewhat less than a factor $\exp(15-\lambda\langle E\rangle_{seq}) = 20$ larger (cf. eqn (A17)), giving possibly 30,000 pseudopromoters in the genome. Taken at face value, these numbers would indicate a significant initiation by RNA polymerase at non-specific sites in the genome. However, this is likely to represent an overestimate as a measure of the number of functional and accessible pseudopromoters.

First, only a fraction (perhaps less than 10%) of all non-specific sites are expected to be accessible for recognition at any one time. Second, there may exist subtle requirements for promoter recognition other than the primary sequence specificity considered in the present calculation; e.g. higher-order (beyond nearest-neighbor) correlations between different base-pairs or contributions from regions surrounding the 30 base-pair site size used in the analysis here. It is also likely that strong pseudopromoters are selected against. However, most of the non-specific activity is expected to derive from pseudopromoters with weak homology since there are so many more of them. It appears less likely that an effective selection will be operating against the large number of weak pseudopromoters; this is corroborated by the fact that Mulligan *et al.* (1984) find almost exactly as many (1396) "promoter-like sequences" in plasmid pBR322 as expected (about 1380) from random occurrence, using their particular definition of a promoter-like sequence. Thus, the expected number of randomly occurring pseudosites as given by equation (27) serves as an interesting reference point for the specificity requirements.

Since the selection parameter, $\lambda$, is equal to unity in the promoter sample, we can use equation (14b) directly to estimate the average activity $\langle k_2 K_B\rangle_{rnd}$ for a random site in the genome:

$$\langle k_2 K_B\rangle_{rnd} \approx (k_2 K_B)_{max} \times \exp(-\langle E\rangle_{seq} - I_{seq})/f(L_{opt}). \quad (28)$$

(The factor $1/f(L_{opt}) \approx 2$ accounts for the different spacer lengths as required by eqn (A22).) Assuming that the free polymerase concentration in the cell is $3\times10^{-8}$ M (McClure, 1985), and using the other data for the promoter sample as above, this gives $\sim$500 $F_A$ initiations per second at random sites in the genome. (It should be noted that this estimate holds even if the maximum activity $(k_2 K_B)_{max} \approx 10^{11}\ \mathrm{M}^{-1}\ \mathrm{s}^{-1}$ is not attainable since the total activity at pseudopromoters is dominated by the weaker ones.) If it is further assumed (arbitrarily) that only 5% of the genome is accessible for RNA polymerase ($F_A = 0{\cdot}05$) and that a random transcript is only $\sim$200 bases long (i.e. the transcript would take approximately 4 s to complete), the total number of polymerase molecules active in random transcription would be $\sim$100. This is about 3% of the total number of actively transcribing polymerase molecules (McClure, 1985). It does not appear likely that a much larger fraction would be allowed, and probably the fraction should be even smaller. This calculation is intended primarily to illustrate the possible consequences of the specificity requirements. It seems clear, however, that if random initiations do occur it is crucial that the transcripts started at these loci be rapidly terminated.

On the basis of primary sequence specificity the promoters do not appear to be overspecified; their numbers are not in large excess over that which would be expected from random appearance in the genome. Instead, efficient discrimination from pseudopromoters may be achieved by keeping the control regions more accessible than the average DNA. For instance, control regions could have sequence characteristics that make them unlikely to be covered by structural proteins. Such secondary sequence specificity (von Hippel & Berg, 1986) could reside in a sequence choice that subtly changes the DNA helix parameter over larger stretches of DNA (Drew & Travers, 1984) or it could reside in a combination of effects from the various recognition sites that make up the control region. Furthermore, a random RNA transcript is not likely to be translated and could therefore be quickly terminated by, for example, *rho*-dependent transcription termination (von Hippel *et al.*, 1984; Platt, 1986). Thus, part of the effective promoter specificity may reside in a close coupling with

ribosome initiation sites. In contrast, operator sites that rely on an equilibrium selection cannot show such kinetic discrimination and may therefore require more overspecification for optimal specificity.

## 4. Evolutionary Selection of Binding Sites

The theory described here relies on the assumption that specific sequences have been positively selected to provide a certain binding affinity or biological activity. Opposing this specific selection is the mutational drift towards randomness. To weigh the importance of certain sequence choices it is necessary to know, as a base-line, what the random base-pair choice is; for this we have simply used the average base-pair composition of the genome. This is a natural assumption, although not necessary. Operationally, the random choice could be represented by the composition of a part of the genome that is under no selection pressure whatever.

While we do not yet understand the significance of the particular distribution of specificity found for the promoter sites shown in Figure 3, it is interesting to discuss some of the factors that can influence and shape such a distribution. Although the specificity must reflect the functional requirement for specific activity at individual sites, it is likely to be further influenced by the particular properties of sequence drift and selection.

The evolutionary constraint will work both on the recognizer protein (affecting the discrimination energies $\varepsilon_{lB}$) and on the binding sequences $\{B_l\}$ actually used. A minimal requirement for effective binding selection in the living cell would be that sequences and discrimination factors are both chosen large enough to reduce the competitive binding to strong pseudosites in the genome to appropriate levels. In this way the investment in protein can be kept low. A larger site size can allow weaker discrimination factors without losing effectiveness in binding selection. This would also permit a much larger variability in the specific binding sequences actually used.

Alternatively, if the discrimination factors are very large, the specific sites could be defined using a minimum number of base-pairs, but would also allow a minimal variability in the binding sequences; large discrimination factors require a very precise protein–DNA interaction that does not permit much variability either in the protein sequence or in the DNA sequences. In a sense, this approach would correspond to a maximization of specificity.

Maximizing specificity by decreasing site size to a minimum and increasing the discrimination factors may lead to some reduction in the investment in protein that is required for a suitable binding level. This might well represent some gain in efficiency and evolutionary fitness. However, this gain is probably not sufficient to counteract totally the continuous drift towards disorder. That is, there are always many more sequences (DNA and protein) that can support weak binding interactions. A balance will be reached when the entropic drift towards smaller discrimination factors and less perfection requires too heavy an investment in protein to permit sufficient binding. The natural fluctuations of protein numbers in the living cell, which can be very large (Berg, 1978), set another limit to how good specificity can usefully be. As discussed previously (von Hippel & Berg, 1986), a regulatory system with too-high specificity would be very sensitive to the removal of even a single protein molecule by fluctuation in protein concentration.

A larger site size requires a larger protein to recognize it. Such enlargements can be achieved by the formation of dimers (or of larger multimers) of the protein. An effective increase in site size can also be achieved by co-operative binding of the same protein to two neighboring binding sites, so that the effective recognition sequence consists of the two sites taken together. Apart from the fact that co-operative binding can have different regulatory sensitivities, there may also be a substantial gain in specificity in such an arrangement; although more protein is required for specific binding, the reduction in the number of competitive pseudosites will be very large, so that the protein "wasted" by non-productive binding can be substantially reduced.

As the whole system grows more complicated, it can also make use of combinations of specific processes and thereby relax the specificity requirements in the individual reactions. Some examples of this possibility were discussed above in connection with the apparent lack of overspecification for the promoter sites. As a corollary of this, it can be expected that more primitive systems have higher requirements for primary sequence specificity.

The large variability observed for the real sites implies that specificity has not been maximized in evolution. This is also corroborated by the fact that the best binding sequences seemingly are not utilized either for the promoter sequences discussed above or for the *lac* operator (Sadler *et al.*, 1983; Simons *et al.*, 1984). In the picture developed above, this is understandable in terms of the fact that, whenever possible, sequence drift would tend towards weaker sites since there are so many more of them. The large variability could reflect the real difficulty of designing a protein with very large discrimination factors, i.e. one with a very precise recognition surface for DNA binding. However, even if specificity could be absolute, there may be advantages to using many small discrimination factors rather than a few strong ones. Thus, the use of some weaker discrimination factors permits a fine tuning or modulation of the binding (or activity) at different specific sites.

However, the observed sequence variability seems to go beyond such requirements for fine tuning. It appears very likely that the choice of discrimination factors (and thereby the permitted

variability in sequences) will tend towards a situation that is most stable in an evolutionary sense. Such a situation will be reached when most mutations, in either the protein or the specific DNA sites, have only a small influence and are not singly lethal. This is consistent with the notion that many small discrimination factors are better than a few strong ones. Furthermore, one would expect a mutationally stable situation to provide many pathways for revertants that can restore binding or activity. This would imply that neither the protein nor the specific sequences used are the best binders, but rather that they incorporate many positions where a mutation could lead to better binding as well as to weaker interactions. Indeed, both repressor mutations (Nelson & Sauer, 1985) and operator mutations (Sadler *et al.*, 1983; Simons *et al.*, 1984) have been found that show increased binding affinities over the wild-type protein molecules and operator sites.

Flexible and imperfect recognition of this sort may also favor independent, rather than co-operative, base-pair interactions, since co-operativity would imply that many base-pair interactions can be lost as a consequence of a single mutational event. Such a flexible recognition interaction would be most stable not only with respect to DNA mutations, but also with respect to translational errors in the recognizer protein molecules. From the level of translational errors observed, one can infer that particular protein molecules should be viewed as members of a family where some entities can have slightly different properties depending on which and how many amino acid substitutions have been incorporated, rather than as identical units (Ehrenberg & Kurland, 1984). Although some amino acid substitutions no doubt will lead to a non-functional protein, most will just result in slightly changed protein properties. Again, stability would argue that the recognizer protein is not designed for maximum specificity. Rather, the optimum will occur where most substitutions have a small effect, with some leading to increased and others to decreased specificity. Thus, the natural tendency will be towards lower specificity and higher disorder, simply because there are many more sequences (DNA and protein) that can fit a lower specificity requirement. While there is little advatage in (and therefore little selection for) maximum specificity, there will be a strong selection against *too* low a specificity.

Thus, rather than maximizing specificity, evolution will tend to *minimize the maximum loss of specificity*. In fact, one can expect the same principle to hold for the design of protein molecules in general if specificity is replaced by specific activity. This agrees with the effects found for various amino acid substitutions in some enzymes (e.g. bacteriophage T4 lysozyme; Tom Alber, Brian Matthews *et al.*, unpublished results) where most substitutions have small effects and some lead to increased activity. Similarly, thermal stability may be a property that is not strongly selected for. Since many more sequences with low stability are expected to exist, the natural tendency would be for proteins to show only the minimal necessary thermal stability. Again this is consistent with the experimental findings (e.g. for T4 lysozyme; John Schellman *et al.*, personal communication) indicating that most amino acid substitutions result in only small increases or decreases in thermal stability.

## 5. Discussion

Our statistical-mechanical selection model provides a physical basis for the sequence analysis of specific DNA sites. It includes the information-theoretic description (Schneider *et al.*, 1986) as a limiting case when only entropy is considered. The theory not only predicts the correlations between promoter activity and "homology score" proposed by Mulligan *et al.* (1984), but also accounts for the observed deviations in terms of the expected statistical uncertainty. Actually, rather than being a "blemish" the observed scatter in the correlation lends further support to the theory.

In essence, the theory presented above consists of two parts. The statistical-mechanical sequence analysis enables us to predict the influence on specificity of individual base-pair choices. This part, which carries a very large statistical uncertainty, can be combined with (or superceded by) actual measurements of base-sequence-dependent changes in activity (or affinity). The second part assumes that the discrimination factors for individual base-pairs are known and calculates the effective specificity in terms of competition from pseudosites, etc. Taken together, the two parts enable us to make quantitative predictions about the specificity of particular DNA sequences, as well as to put statistical measures (notably sequence information) into the context of the regulatory requirements of the living cell.

We have derived from first principles a relation between DNA sequence variability in the binding sites and the binding affinity (or activity) for the particular protein that recognizes these sites. As discussed above, neither of the two basic assumptions (the equiprobability for all sequences with the same specific affinity (or activity) and the independence of individual base-pair contributions) can be strictly true in general. In the analysis of the promoter sample we identified some deviations from both assumptions. However, these are not dominant effects and the results of the analysis (being within or close to the expected statistical uncertainties) provide no justification for revision or refinements of the basic assumptions at this time. As more sequence data accumulate and as more binding (or activity) constants are measured, some such refinements will no doubt be required, thus providing more information on physical and evolutionary constraints for regulatory site function.

In essense a sequence analysis of this sort mixes all of the sequences, and then extracts binding information from the patterns of base-pair utilization frequencies. Although this procedure obviously must lead to large statistical uncertainties, one advantage is that sequence requirements that are not shared by many of the sites (e.g. for effector binding) will be averaged out. In contrast, when many sites in the sample share constraints not related to primary protein recognition, the equiprobability assumption is invalid and the results will be skewed, signalling the need to look for additional constraints.

Similarly, the interpretations could be skewed if many sites in the sample are derived from the same basic sequence, as would be the case if they had recently evolved from some common ancestor. While refinements such as introducing higher-order base-pair correlations or using a different weighting scheme for the selection constraint can be introduced, the approach presented here should provide the essentials of what one can do with sequence data alone. The usefulness of this analysis can only be judged by its success in predicting binding or activity for specified sequences. Again, large discrepancies would be an indication of selection constraints other than binding affinity, and could possibly be used to help identify such additional constraints.

Although the evolutionary selection of binding sites take place *in vivo*, one expects the detailed correlations between sequence and binding (or activity) to show up *in vitro* where the differential influence of effectors other than DNA sequence can be kept to a minimum. The main requirement is that the property selected for *in vivo* be the same as that studied *in vitro*. To the extent that sequence is important for specificity, some selection constraint will be operating even if *in vivo* activities are strongly influenced by other effects as well; such constraints will then show up in the relations for *in vitro* activity, where sequence can be made to play an even more dominant role.

The theory developed above also puts in context the various levels of selection that determine the observed sequences. The *binding selection* by the protein is based directly on the discrimination factors. The *evolutionary selection* of specific sequences is constrained by the binding selection, but could be biased in various ways. Also sample selection of the sites that have been identified and sequenced may be biased. In principle, we cannot distinguish these sources of bias. However, the theory works for an average selection constraint and is not much influenced by variations or bias around this average.

Although developed in terms of binding affinity (or activity for the promoters), the theory is valid when selection is based on any property for which the contributions from individual base-pairs can be considered additive. The theory should be applicable not only to protein–DNA specificity, but also to protein–RNA specificity and possibly to interactions between nucleic acids as well (e.g. the ribosome binding sites; Gold *et al.*, 1981). However, the application to these other systems may be less useful, since interactions involving single-stranded nucleic acids may be less linearly constrained, allowing effective "rearrangements" of the sequence simply by the extrusion of non-complementary sections of the RNA (or the single-stranded DNA) from the binding interaction.

This research has been supported by United States Public Health Service, research grants GM-15792 and GM-29158 (P.H.v.H.), and by salary support from the Swedish Natural Science Research Council (to O.G.B.). We are grateful to Drs Gary Stormo and Larry Gold for useful discussions and correspondence, and to Drs Schneider, Stormo, Gold and Ehrenfeucht for sending us their manuscript prior to publication. We also thank Ms Jean Parker for her skill and patience in typing and retyping the many versions of the manuscript.

## References

Berg, O. G. (1978). *J. Theoret. Biol.* **71**, 587–603.
Bremer, H. & Dalbow, D. G. (1975). *Biochem. J.* **150**, 9–12.
Crooks, J. H., Ullman, M., Zoller, M. & Levy, S. B. (1983). *Plasmid,* **10**, 66–72.
de Crombrugghe, B., Busby, S. & Buc, H. (1984). *Science,* **224**, 831–838.
Dickerson, R. E. (1983). *J. Mol. Biol.* **166**, 419–441.
Drew, H. R. & Travers, A. A. (1984). *Cell,* **37**, 491–502.
Ehrenberg, M. & Kurland, C. G. (1984). *Quart. Rev. Biophys.* **17**, 45–82.
Galas, D. J., Eggert, M. & Waterman, M. S. (1985). *J. Mol. Biol.* **186**, 117–128.
Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Swebilius Singer, B. & Stormo, G. (1981). *Annu. Rev. Microbiol.* **35**, 365–403.
Gurney, R. W. (1949). *Introduction to Statistical Mechanics*, McGraw-Hill Book Company, New York.
Harr, R., Haggstrom, M. & Gustafsson, P. (1983). *Nucl. Acids Res.* **11**, 2943–2957.
Hawley, D. K. & McClure, W. R. (1983). *Nucl. Acids Res.* **11**, 2237–2255.
Jobe, A., Sadler, J. R. & Bourgeois, S. (1974). *J. Mol. Biol.* **85**, 231–248.
Lamond, A. I. (1985). *Trends Biochem. Sci.* **10**, 271–274.
McClure, W. R. (1985). *Annu. Rev. Biochem.* **54**, 171–204.
Mossing, M. C. & Record, M. T., Jr (1985). *J. Mol. Biol.* **186**, 295–305.
Mulligan, M. E. & McClure, W. R. (1986). *Nucl. Acids Res.* **14**, 109–126.
Mulligan, M. E., Hawley, D. K., Entriken, R. & McClure, W. R. (1984). *Nucl. Acids Res.* **12**, 789–800.
Nelson, H. C. M. & Sauer, R. T. (1985). *Cell,* **42**, 549–558.
Ohlendorf, D. H., Anderson, W. F., Fisher, R. G., Takeda, Y. & Matthews, B. W. (1982). *Nature (London),* **298**, 718–723.
Platt, T. (1986). *Annu. Rev. Biochem.* **55**, 339–372.
Rossi, J. J., Soberon, X., Marumoto, Y., McMahon, J. & Itakura, K. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 3203–3207.
Sadler, J. R., Sasmor, H. & Betz, J. L. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 6785–6789.
Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. (1986). *J. Mol. Biol.* **188**, 415–431.

Simons, A., Tils, D., von Wilcken-Bergmann, B. & Muller-Hill, B. (1984). *Proc. Nat. Acad. Sci., U.S.A.* **81**, 1624–1628.
Staden, R. (1984). *Nucl. Acids Res.* **12**, 505–519.
von Hippel, P. H. (1979). In *Biological Regulation and Development* (Goldberger, R. F., ed.), vol. 1, pp. 279–347, Plenum, New York.
von Hippel, P. H. & Berg, O. G. (1986). *Proc. Nat. Acad. Sci., U.S.A.* **83**, 1608–1612.
von Hippel, P. H., Bear, D. G., Morgan, W. D. & McSwiggen, J. A. (1984). *Annu. Rev. Biochem.* **53**, 389–446.

*Edited by C. R. Cantor*

APPENDIX

# Statistical Ensembles for Sequence Variability

**Otto G. Berg**

We shall start by considering the set of all potential sites; i.e. all sequences that could possibly work as specific binding sites for a particular recognizer protein. Whether a sequence is a potential site or not is determined by its binding affinity. For simplicity, it is assumed that each base-pair in the binding sequence contributes independently to the binding free energy. This independence assumption implies that each possible base-pair at each position in the binding sequence can be assigned a unique *discrimination energy* $\varepsilon > 0$, defined as the difference between the binding free energies for a sequence with the best binding (cognate) base-pair at this position and the one with the actual base-pair under consideration. It also implies that there exists a best binding sequence (with maximum binding affinity) to which all other sequences can be uniquely related.

## (a) *The cut-off distribution*

### (i) *Base-pair choice*

First we shall assume that there exists a threshold in binding affinity so that sites with weaker binding are not acceptable as specific sites. In the simplest case, each non-cognate base-pair in the sequence decreases the binding affinity by the same amount $\varepsilon$ (this is here defined as a free-energy difference in units of $kT$). If the total binding affinity can be reduced from the maximum affinity by at most $E_c$ if a site is to be potentially useful as a specific site, we can easily derive the variability in the sequence distribution of all potential specific sites. Out of a total of $4^s$ different sites of length $s$, the number of potential binding sites with $j$ base substitutions is given by (cf. eqn (6) of von Hippel & Berg, 1986):

$$N_j = \begin{cases} \binom{s}{j}3j; & \text{for } j \le E_c/\varepsilon \\ 0; & \text{for } j > E_c/\varepsilon \end{cases}, \qquad \text{(A1)}$$

if base-pairs are *a priori* equiprobable. Thus, the total number of potential binding sites of length $s$ is given by:

$$W_s(E_c) = \sum_0^J \binom{s}{j}3^j \approx \binom{s}{j}3^J \approx \frac{(3s/J-3)^J}{[2\pi J(1-J/s)]^{\frac{1}{2}}(1-J/s)^s}, \qquad \text{(A2)}$$

where $J = \mathrm{Int}(E_c/\varepsilon)$ is the maximum number of substitutions that leave the binding affinity within $E_c$ from the maximum one. The first approximation replaces the sum in equation (A2) with its maximum term, which is reasonable as $N_j$ of equation (A1) increases rapidly with $j$. The second approximation step uses Stirling's formula $[\ln(n!) \approx n(\ln n - 1) + (1/2)\ln(2\pi n)]$ for the factorials in the binomial coefficient and is given here for later use below.

Among the $W_s(E_c)$ potential binding sites,

$$3\sum_{j=0}^{J-1} \binom{s-1}{j}3^j$$

will have a base-pair substitution at any particular position in the site. Thus, the fraction of potential sites that have a substitution at a certain position $l$ in the site is:

$$b_l = 3\sum_{j=0}^{J-1} \binom{s-1}{j}3^j / W_s(E_c) \approx J/s \approx \{(E_c/\varepsilon) - (1/2)\}/s, \qquad \text{(A3a)}$$

where the first approximation is the same as in equation (A2), i.e. we replace the sums by their respective maximum term. The second approximation replaces the integer function $J = \mathrm{Int}(E_c/\varepsilon)$ with its continuous approximation $J \approx E_c/\varepsilon - 1/2$. In this simple case, the frequencies of base utilization among the potential sites is:

$$f_{l0} = 1 - b_l = 1 - J/s, \qquad \text{(A3b)}$$

for the cognate base-pair ($B = 0$), and:

$$f_{lB} = b_l/3 = J/3s, \qquad \text{(A3c)}$$

for each of the three non-cognate ones ($B = 1, 2, 3$) at each position $l$.