Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 7501–7505, August 1992
Evolution

The evolutionary selection of DNA base pairs in gene-regulatory binding sites

(gene regulation/protein–DNA binding/mutation-selection balance/protein burden/evolution)

OTTO G. BERG

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

Communicated by Peter H. von Hippel, May 8, 1992

ABSTRACT **The DNA base-pair sequences that serve as gene-regulatory sites have been selected during evolution to provide an appropriate functional binding for a particular protein. In most cases, the function depends on the binding probability, which can be influenced both by the binding strength and by the abundance of the protein in the cell. As a consequence, the same function can be achieved with strong binding sites and a small amount of protein as with weak binding sites and a large amount of protein. However, increasing the protein burden will decrease the growth rate of the cells, even when all functions remain the same. Thus, for maximal growth, the protein levels should be as low as possible and the binding correspondingly strong. On the other hand, sequences with a weaker binding can be formed in many more ways and are, therefore, more probable, and random mutations are more likely to produce them. Thus, the selection pressure against an increased protein burden can be balanced against the random mutational drift in the recognition sequences, thereby tying together the statistics of base-pair choice, the binding strength, and the protein burden. In terms of this model, the selection pressure can be estimated from the properties of a gene-regulatory protein and its recognition sites. A key feature is the mutational randomization pressure that appears as a fundamental force shaping the optimal solutions that provide maximal growth. The model is tested on a number of gene-regulatory systems in *Escherichia coli*. The same principles should hold for all proteins for which overall activity in the cell is proportional to abundance; then the selective pressure to increase the efficiency of an individual protein cannot be larger than the selective pressure to decrease the total protein burden.**

Many microbial organisms live under intense selection pressure, where the primary competitive advantage, at least sometimes, is given by a high growth rate. This situation leads to the hypothesis that the evolutionary selection of these organisms has optimized their cellular components to produce a maximal growth rate (1, 2). However, this picture cannot be complete without also considering the difficulty of achieving the optimal situation. Opposing the tendency toward ever better solutions will be the random mutations that are always more likely to reduce fitness than to increase it. The closer the system is to the optimal situation, the more likely are the mutations to reduce fitness. This situation is particularly acute for properties under intense randomization pressure. Thus, it has been argued that gene-regulatory DNA sites (3) and the use of nonoptimal codons (4) should be viewed in this way; there are always so many more ways of achieving nonoptimal DNA sequences that their presence can never be totally selected away. Instead, the organism must compromise some of its potential growth-rate advantage for the much higher likelihood of nonoptimal solutions.

Gene-regulatory proteins can bind to a number of functional recognition sites in the genome of an organism. These sites exhibit a large variability in their DNA sequences. Based on the assumption that each sequence has been selected to provide some appropriate binding, the base-pair choices can be correlated with the binding strength of any particular sequence. In many cases—e.g., for repressor and activator proteins—the function is directly proportional to the probability that a certain site is occupied by the protein. This probability, in turn, can be influenced either by the binding constant of the site or, through mass action, by the concentration of the protein in the cell. Thus, evolution can be expected, in some way, to have balanced the cost of increased protein levels (e.g., through the protein burden) against some cost or difficulty of increasing binding strength at the functional sites.

The first relationship to be considered below is that between the statistics of base-pair choice and binding strength at the specific sites. This consideration leads to a measure for the effective randomization pressure on the binding strength of recognition sequences. The second step is to introduce a standard population-dynamic selection model that describes the growth and dominance of favorable phenotypes. This model includes both the randomization pressure and the growth reduction due to an increased protein burden. The most likely variants to be selected are those where the growth reduction due to increased levels of gene-regulatory protein is balanced against the higher likelihood for weaker binding sites. In this way it becomes possible to derive a relationship between the selection of base pairs in the recognition sites, the investment in protein, and the evolutionary selection pressure on the organism.

Base Pair Statistics and Binding

Previously, Berg and von Hippel (3, 5–9) have studied the relationship between the statistics of base-pair choice and the functional strength of the recognition sites for certain gene-regulatory proteins. From the statistics of base-pair occurrence in the set of natural recognition sites for a gene-regulatory protein, r, one can define a *dissimilarity index*, D, for any base-pair sequence $\{B_1B_2B_3 \ldots B_s\}$:

$$D = \sum_{j=1}^{s} \ln\left[\frac{n_{j0} + 0.5}{n_{jBj} + 0.5}\right] \qquad [1a]$$

$$\sigma_D = \sqrt{\sum_{j=1}^{s}\left[\frac{1}{n_{jBj} + 0.5} + \frac{1}{n_{j0} + 0.5}\right]}, \qquad [1b]$$

where n_{jBj} is the number of occurrences in the natural sites of a particular base pair B_j (AT, TA, CG, or GC) at position j; n_{j0} is the number of occurrences of the most common base pair (the consensus base pair) at this position. The sums are

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

taken over all positions $j = 1, 2, \ldots, s$ (where s is the site size), except those that carry a consensus base pair. σ_D is the expected statistical uncertainty (SE) in the numerical value of the index. The dissimilarity index (in previous communications denoted λE) is a measure of the statistical sequence difference of the studied site from the consensus sequence for the natural sites. Similar indexes—or homology scores—have been defined and used by others (10–12), but the one given in Eq. **1** has the advantage that it is expected to be directly related to the functional strength of a recognition sequence.

From theoretical analysis (3, 7, 9) and applications to experimental data (3, 5, 8), it has been demonstrated that the correlation between the statistics of base-pair choice and the binding constant K of a particular site can be expressed as:

$$\ln(K) = \ln(K_0) - (1/\lambda_r)D, \quad [2]$$

where K_0 is the binding constant to the consensus sequence, and the parameter λ_r is determined from the slope of the regression line. From this expression $(1/\lambda_r)D$ corresponds to the reduction in binding free energy (in units of kT) due to the presence of nonconsensus base pairs in the sequence. In this way λ_r is primarily a correlation coefficient that expresses the coupling between statistics and function.

An interpretation of the significance of the parameter λ_r derives from a consideration of the probability density $P_{rnd}(E)$ for random base-pair sequences to provide a certain binding free energy E (in units of kT) for the protein r. It can be shown (3, 7, 9) that:

$$\lambda_r = [d(\ln(P_{rnd}))/dE]_{E=E_s}, \quad [3a]$$

so that λ_r expresses the rate of increase of this probability density as binding strength decreases (increasing E) from the level E_s for the average natural site:

$$P_{rnd}(E) \propto \exp(\lambda_r E) \qquad (\text{around } E \approx E_s). \quad [3b]$$

Thus, λ_r expresses the asymmetry in the distribution for random sites in the neighborhood of the specific ones; for large λ_r values there will be many more sequences with weaker binding than the specific ones. One finds (3, 7) that λ_r increases strongly for more constrained sites (i.e., for small E_s) and approaches zero when the average binding strength of the specific sites approaches random binding. In this way, λ_r can be seen as a measure of the *randomization pressure* on the binding strength of the recognition sites because mutations would be more likely to drive a site toward weaker binding for which there are many more sequence possibilities available, particularly for large λ_r values. As a consequence, it can be expected that λ_r in some sense also represents a selection pressure that balances this randomization tendency. To explore the extent to which this conjecture could be true, one needs a model for the evolutionary selection of the system properties of the organism.

Selection

The ratio of the probabilities of selection for two genetic variants that differ in one mutation and in their relative growth rates by an amount s_i can be expressed as (14, 15):

$$P_I/P_0 = \exp(2N_e s_i), \quad [4]$$

where N_e is the effective population size. The *selective advantage* s_i is negative if the variant i has slower growth. Thus, the ratio for the probabilities of occurrence of any two genetic variants (differing only at one position and that differ in relative growth rates by s) is given by $\exp(2N_e s)$.

Consider a number of different independent positions where mutations produce additive changes in growth rate. Then the probability for a set of such mutations ($ijk \ldots$) at positions 123 . . . in the genome is

$$P_{ijk\ldots} = P_{000\ldots} \exp[2N_e(s_{i1} + s_{j2} + s_{k3} + \ldots)]. \quad [5]$$

As a consequence, the ratio between the probability for all possible variants with the same phenotype I (same s and same functions) that can be achieved in g_I different ways and some basic variant that can be achieved in g_0 ways will be

$$P_I/P_0 = (g_I/g_0)\exp(2N_e s). \quad [6]$$

In this way the *degeneracy factors* g must enter the probability expressions when the same function can be achieved in different ways. These factors express the tendency for random mutations to push toward more likely situations. If there is no growth rate difference ($s = 0$), Eq. **6** simply corresponds to random choice.

Let us compare the basic variant, 0, with one, I, where the binding free energy for all the n_r regulatory sites for a protein r has been shifted by δE. Then from Eq. **3**, the ratio of the degeneracy factors would be

$$g_I/g_0 = \exp(n_r \lambda_r \delta E). \quad [7]$$

Assuming that the difference in relative growth rate is δs, the ratio of the probabilities of occurrence for the two variants will be

$$P_I/P_0 = \exp(n_r \lambda_r \delta E)\exp(2N_e \delta s). \quad [8]$$

Considering all such variations, δE, one finds that the maximum probability is for those that have a relative growth-rate change:

$$\delta s = -n_r \lambda_r \delta E/2N_e. \quad [9]$$

This is the *maximum likelihood*: The binding strength of the regulatory sites are most likely to be in a region where small changes lead to a growth rate change as given by Eq. **9**. Of all the possible combinations, this is the one that has the largest probability of having been selected. If there is no difference in the degeneracy factors (i.e., $\lambda_r = 0$), then Eq. **9** would correspond to growth maximization where $\delta s = 0$.

Protein Burden and Gene Regulation

A small change δM_p in the protein burden (for instance by a nonfunctional protein) should lead to a proportional change δs in the relative growth rate constant:

$$\delta s = -\mu \delta M_p/2N_e, \quad [10]$$

where the parameter μ has been defined, such that $\mu/2N_e$ is the proportionality constant. M_p is the total amount of amino acids invested in protein in the cell. All things being functionally equal (by assumption), the variant with the smaller protein investment should have *some* growth advantage, and Eq. **10** can be considered as the first term in a general series expansion. Then using Eq. **6**, one finds that two genetic variants that differ only in their protein investment (but not in any function or activity related to this protein) will occur in the ratio

$$P_I/P_0 = (g_I/g_0)\exp(-\mu \delta M_p). \quad [11]$$

This relation shows that μ as defined by Eq. **10** can be interpreted as the selection pressure against small increases in the protein burden.

Consider a gene-regulatory protein r that is built up from N_r amino acids and is present in C_r copies in the cell. Assume further that it has n_r functional recognition sites in the genome, each with a binding constant $K_i = K_0\exp(-E_i)$. The binding occupancy at each site i ($i = 1, 2, \ldots, n_r$) is

$$\theta_i = \frac{K_iC_r}{1 + K_iC_r} = \frac{C_rK_0\exp(-E_i)}{1 + C_rK_0\exp(-E_i)}. \quad [12]$$

This equation assumes that the law of mass action applies within the cell; even if the chemical activity of a gene-regulatory protein is not equal to the concentration, it should be proportional. In this way nonideal intracellular solution effects can be incorporated in the absolute values of the binding constants. These absolute values are unknown and irrelevant for the following discussions, which rely only on the relative binding constants.

If the protein amount is changed by δC_r, its functional activity (binding occupancy at each of the n_r natural sites in the genome) will change in proportion. However, this change could be compensated by decreasing the binding free energy at each of the n_r sites by the same amount $\delta E = \delta C_r/C_r$. Such a compensatory change would keep the binding relation $C_r\exp(-E)$ invariant [i.e., $C_r\exp(-E) = (C_r + \delta C_r)\exp(-E - \delta E)$]. Thus, we can consider two functionally equivalent genetic variants, one with a higher level of protein r and weaker binding constants and the other with a lower level of the protein and stronger binding constants.

The growth rates of the two functionally equivalent variants will differ only due to their difference in protein burden, which is $\delta M_p = N_r\delta C_r$. Thus the ratio of the probabilities for the two variants will be given by Eq. **11** (using Eq. **7** for the degeneracy factors):

$$P_I/P_0 = \exp(n_r\lambda_r\delta E - \mu N_r\delta C_r). \quad [13]$$

By considering all such functionally equivalent variants for which $\delta E = \delta C_r/C_r$, one finds that the maximum probability is given by those variants for which

$$n_r\lambda_r = \mu N_rC_r. \quad [14]$$

In this way, the selective growth advantage of one variant is balanced against the larger number of possibilities for the other. The parameter choices most likely to occur are those that have the maximal probability under the constraints that produce the observed functional properties of the organism. Eq. **14** provides the fundamental relation between the statistics of base-pair choice (through $\lambda_r n_r$) in a set of recognition sites and the investment (N_rC_r) in the relevant regulatory protein coupled via the selection pressure μ as defined by Eq. **10**.

Consequences

Eq. **14** relates the properties of the binding sites with the investment in gene-regulatory protein through the selection pressure μ. When the exact binding relations (Eq. **12**) are considered, which involve the free and not the total protein amounts, one finds that the quantity C_r that enters Eq. **14** would be the *excess* amount of the protein (i.e., the amount not bound at functional sites). It does not matter whether some or most of this excess is bound at nonspecific DNA sites because the concentration of free protein is expected to be proportional to the concentration of the excess; then Eq. **14** holds with C_r being the excess amount. In most cases the difference between the total and the excess amounts is not very large. From Eq. **14** one expects that the ratio $\lambda_r n_r/N_rC_r = \mu$ should be an invariant for all gene-regulatory systems of an organism.

The data for some gene-regulatory systems from *Escherichia coli* have been collected from the literature and listed in Table 1 (refs. are given in the first footnote). The main problem is to find reasonable estimates for the number of proteins per cell; only cases where the uncertainty in this number is a factor of 5 or less have been included. From Table 1 one finds that μ calculated from Eq. **14** is fairly constant around $\mu \cong 10^{-4}$ in *E. coli*. The invariance holds reasonably well, considering the large uncertainties in many of the numbers involved. A possible exception is the *arg* repressor (r = 4 in Table 1), which, if it is, indeed, present in only 40 copies in the cell, would seem to have a stronger selection against increasing this number than expected. [Recent results (13) in *Salmonella typhimurium*, however, indicate that the *arg* repressor binds simultaneously to two sites. This binding would reduce by half the number of sites, n_r, and thereby bring the corresponding μ value more in line with those of other systems.]

The extent to which the growth rate is decreased by an increased protein burden has been a matter of some controversy (20, 21). Naively one might expect that the relative decrease in growth rate is equal to the relative increase in protein investment:

$$\delta_s = \delta M_p/M_p. \quad [15]$$

Using a theoretical model for growth maximization (2), one can show (O.G.B., unpublished work) that this is, indeed, a lower limit for the optimal relation. Also experimentally this relationship could hold, although interpretations of the experimental data have been difficult (21). Using the estimate for δ_s from Eq. **15** above and Eqs. **10** and **14**, one finds that the effective population size would be given by

$$N_e = \lambda_r n_r/2f_r, \quad [16]$$

where $f_r = N_rC_r/M_p$ is the fraction of the total protein mass that is invested in the excess of the gene-regulatory protein r under consideration. This is also listed in Table 1. The resulting estimate for N_e between 2×10^4 and 8×10^4 would be unreasonably low (14) if it were an estimate of a real population size. However, in this application, N_e is best

Table 1. Testing of model on gene-regulatory systems from *E. coli*

r*	n_r	λ_r†	N_r	C_r‡	$\mu \times 10^5$§	f_r¶	$2N_e \times 10^{-5}$‖
1	1000	1.0	4000	1200–6000	20–4	0.007–0.01	1.4–1
2	17	1.3	2 × 202	650	8	0.00013	1.7
3	100	0.8	2 × 209	3000	6	0.0013	0.6
4	16	2.1	6 × 156	40–200	90–20		
5	1		4 × 347	10	7	0.000017	0.6
6	3		2 × 115	30–150	40–9	0.00008	0.4

n_r, Number of recognition sites in genome; N_r, number of amino acids in the active multimer of the protein.

*Type of regulatory protein: r = 1, RNA polymerase (3, 16); r = 2, LexA protein (8, 17); r = 3, cAMP receptor protein (5); r = 4, *arg* repressor (8, 18); r = 5, *lac* repressor; r = 6, *trp* repressor (19).

†λ_r values estimated from the statistics–activity correlation given in Eq. **2**.

‡Excess number of protein molecules in the cell; a range of numbers refer to different growth conditions for RNA polymerase and *trp* repressor and to experimental uncertainty in the other cases. Presumably, the uncertainty is much smaller for the cases where no range is given.

§Proposed invariant ratio μ is calculated from Eq. **14**. In the three cases where λ_r is unknown, $\lambda_r = 1$ has been used.

¶Fraction of the total protein mass invested in protein r, $f_r = N_rC_r/M_p$; only cases where this fraction has been explicitly given in the literature are listed.

‖$2N_e$ from Eq. **16**, assuming that $\delta s/\delta M_p = -1/M_p$.

considered as a parameter describing the intragenomic sequence variability in individual cells, or clones, rather than the variability between different clones. It is noteworthy that a study (22) of the codon bias in *E. coli* also leads to a similarly small value of N_e.

The relationship above was derived from a consideration of the most likely variants with given functional properties. It is also interesting to consider what will happen to the growth rate of the cells if there is a mutational change in the level of some gene-regulatory protein. A small increase δC_r in such a level will lead to a change in growth rate both due to the cost of the increased protein burden and due to the functional change from the increased occupancy levels at the regulatory sites. Using Eqs. **9**, **10**, and **14**, one finds that these two effects cancel, so that the total change in the growth rate is zero. Thus, in this model the growth rate is maximized with respect to level of the gene-regulatory protein but is not maximized with respect to strength of the binding sites. A small increase in all binding constants combined with a small decrease in protein amount would lead to a higher growth rate. The maximum likelihood applies to the system parameters, which—like the DNA base-pair choice—is under strong randomization pressure and corresponds to growth-rate maximization for the parameters that are not under strong randomization pressure. (Actually, in the general picture there would be a small contribution to the level of gene-regulatory protein also from the randomization pressure, for instance, on the promoter of its gene.)

As an example, where the repressor level could be optimized in this way to produce a maximal growth rate, let us consider an operon where a repressor controls the production of a number of enzymes. Under repressed conditions, the concentration (the basal level) of enzymes to a first approximation is expected to be inversely proportional to the concentration of repressor in the cell. Consequently, a small decrease in repressor level will lead to a small increase in level of enzymes. It can easily be shown that these two changes will cancel and keep the total protein burden constant if the mass invested in the basal level of enzymes is equal to the mass invested in repressor (this is actually approximately the case for the *lac* operon in *E. coli*). Thus, if these enzymes and the repressor have no other activities in the cell, a small change in repressor level will leave the protein burden and, therefore, also the growth rate invariant. In this case the system would be growth maximized with respect to the repressor level. As a consequence, the growth rate would also be insensitive to small statistical fluctuations in the repressor numbers. Conversely, if these investments are very different, it would imply that the repressor or the enzymes have other functional effects in the cell.

The implications of these calculations are similar to those for the mutation-selection balance that has been proposed for the nonrandom use of synonymous codons (4, 14, 15, 22, 23). In this case the relationship between base-pair choice and function is not as straight-forward as in the case of the recognition sites described here, but the basic result is the same: when selection is not sufficiently strong, the randomization pressure will lead to nonoptimal choices of base pairs. An optimal solution for the system parameters can be found only in the context of balancing the randomization pressure.

One can expect that similar considerations will apply to the function of enzymes; there should be many more amino acid sequences providing a lower activity than a maximal one. Thus, there may be a strong randomization pressure also on these sequences, so that the selection pressure is not sufficiently strong to push the constructions to their limits. Hartl and coworkers (24) have shown how enzymes evolving toward higher kinetic efficiency will reach a point where further improvements will lead to such small increases in fitness that they will be effectively neutral. To this picture we should add the asymmetry in the mutational pressure: the closer to perfection an enzyme is, the stronger the randomization pressure must be. From the actual peak, all paths must go downhill. The randomization pressure on the functional protein sequences, however, is not easily described because there is no straight-forward relationship between sequence choice and functional activity. For many enzymes, it could also be possible to get the same overall activity by increasing their amounts rather than the functional efficiency per enzyme. In such cases the selective pressure to increase efficiency cannot be larger than the pressure to decrease the total protein burden. Thus, it may, in fact, be possible to improve significantly on Nature's designs in many cases.

Discussion

The results were derived from the selection model that leads to Eq. **6**. One major problem is that this result requires some equilibrium in the mutational selection, a mutation-selection balance. This is not likely to be strictly accomplished in a global sense because the system will never have time to explore all mutational variations. Thus, Eqs. **4–6** must be considered locally for variations around some variant that may be the accidental result of the evolutionary history of the organism.

The results are also based on the assumption that a change in the amount of a gene-regulatory protein, while keeping the binding probability at the specific sites constant, will only affect the fitness of the organism through the protein burden. This assumption is obviously an oversimplification. For instance, a large protein excess will lead to faster binding at a site, and this could also be of functional importance. Similarly, the cost of increasing the amount of a gene-regulatory protein may not be only in the increased protein burden; an excess amount could bind at the wrong places in the genome and interfere with the regulation of other genes. Thus, the excess levels of protein could be associated with some advantage and/or cost unrelated to the protein burden. Nevertheless, the binding probability at the functional sites is probably the main functional requirement. Some of the least-abundant proteins (e.g., the *lac* repressor) could already have reached levels where random fluctuations in their numbers could become detrimental if the levels were further reduced (25, 26). However, this constraint should not be a problem for most proteins listed in Table 1.

The basic assumption is the maximum likelihood expressed through Eqs. **13** and **14**, which simply requires that the system parameters are the most probable ones that can give rise to a certain function. In this case the system is functionally constrained such that the required levels of the gene products are strongly selected for and not subject to genetic drift; the levels of gene-regulatory protein and the strength of the recognition sites can be chosen in many ways to produce these gene-product levels. The most likely situations to occur are those where selection and randomization pressures balance each other.

The model has no free parameters. Both $\delta s/\delta M_p$ and N_e should be independently determinable. Application of this selection model to some gene-regulatory sites in *E. coli* gives $\mu = -2N_e(\delta s/\delta M_p) \cong 10^{-4}$; as discussed above, this implies that the effective population size is much smaller for this situation than usually assumed. Alternatively, the randomization pressure on the DNA sites is not sufficient to account for the excess protein numbers that could be set at a high level due to some other requirements. However, the approximate proportionality between the number of sites and the mass of gene-regulatory protein, as displayed by the invariance of μ in Table 1, is the predicted result from the model. It would be more difficult to explain were it not, in some way, based on a consideration of the protein burden; alternative consider-

ations based on the function (e.g., association rates or interference at the wrong sites) would be likely to involve the concentration and not the mass of the protein. More and better data, primarily on the protein abundances, are required to substantiate this selection model. If it holds, the model opens up the possibility of using the results for one gene-regulatory protein to estimate properties of another or even comparing properties of different organisms.

Although many of the details of this model are undoubtedly oversimplified, these basic principles should hold also for proteins other than the gene-regulatory ones: (*i*) The mutational pressure on the sequences (DNA or protein) toward lower functional activity increases the closer a system is to maximal activity. (*ii*) The most likely situation to occur is the one where randomization pressure and selection pressure balance each other. (*iii*) For proteins for which overall activity in the cell is proportional to abundance (mass action), the selective pressure on the functional activity per protein will be limited by the selective pressure on the total protein burden.

Through an expansion of the growth-maximization principle to include the randomization pressure it has been possible to tie together the statistics of base-pair choice with the amounts invested in gene-regulatory proteins and the selection pressure on the protein burden for the organism. This expansion adds another dimension to the description of functional recognition sites. The particularly simple combinatorics of the linear DNA sequences that constitute the recognition sites allowed quantitation of the expected relationships.

I thank Pete von Hippel for his support, Mike Bulmer for discussions and for communicating results before publication, and Måns Ehrenberg and Chuck Kurland for their stimulating challenge to many of the ideas presented in this work. This research has been supported by The Swedish Natural Science Research Council and in its initial stages also by U.S. Public Health Service research grants (to P. H. von Hippel).

1. Maaloe, O. (1979) in *Biological Regulation and Development*, ed. Goldberger, R. F. (Plenum, New York), Vol. 1, pp. 487–542.
2. Ehrenberg, M. & Kurland, C. G. (1984) *Q. Rev. Biophys.* **17**, 45–82.
3. Berg, O. G. & von Hippel, P. H. (1987) *J. Mol. Biol.* **193**, 723–750.
4. Bulmer, M. (1987) *Nature (London)* **325**, 728–730.
5. Berg, O. G. & von Hippel, P. H. (1988) *J. Mol. Biol.* **200**, 709–723.
6. Berg, O. G. & von Hippel, P. H. (1988) *Trends Biochem. Sci.* **13**, 207–211.
7. Berg, O. G. (1988) *J. Biomol. Struct. Dynamics* **6**, 275–297.
8. Berg, O. G. (1988) *Nucleic Acids Res.* **16**, 5089–5105.
9. Berg, O. G. (1990) *Biomed. Biochim. Acta* **49**, 963–975.
10. Harr, R., Häggström, M. & Gustafsson, P. (1983) *Nucleic Acids Res.* **11**, 2943–2957.
11. Staden, R. (1984) *Nucleic Acids Res.* **12**, 505–519.
12. Mulligan, M. E., Hawley, D. K., Entriken, R. & McClure, W. R. (1984) *Nucleic Acids Res.* **12**, 789–799.
13. Lu, C.-D., Houghton, J. E. & Abdelal, A. T. (1992) *J. Mol. Biol.* **225**, 11–24.
14. Li, W.-H. (1987) *J. Mol. Evol.* **24**, 337–345.
15. Shields, D. C. (1990) *J. Mol. Evol.* **31**, 71–80.
16. Shepard, N. S., Churchward, G. & Bremer, H. (1980) *J. Bacteriol.* **141**, 1098–1108.
17. Sassanfar, M. & Roberts, J. W. (1990) *J. Mol. Biol.* **212**, 79–96.
18. Cunin, R., Glansdorff, N., Pirard, A. & Stalon, V. (1986) *Microbiol. Rev.* **50**, 314–352.
19. Kelley, R. L. & Yanofsky, C. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 3120–3124.
20. Diamond, J. M. (1986) *Nature (London)* **321**, 565–566.
21. Koch, A. L. (1983) *J. Mol. Evol.* **19**, 455–462.
22. Bulmer, M. (1991) *Genetics* **129**, 897–907.
23. Sharp, P. M. & Li, W.-H. (1986) *J. Mol. Evol.* **24**, 28–38.
24. Hartl, D. L., Dykhuizen, D. E. & Dean, A. M. (1985) *Genetics* **111**, 655–674.
25. Berg, O. G. (1978) *J. Theor. Biol.* **71**, 587–603.
26. von Hippel, P. H. & Berg, O. G. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 1608–1612.