Reviews

Selection of DNA binding sites by regulatory proteins

Otto G. Berg and Peter H. von Hippel

As a consequence of progress in several fields, we are developing an understanding of the structural, thermodynamic and statistical rules that govern the binding of regulatory proteins to functional sites on the DNA genome. The current state of our understanding of these rules, and some approaches to the evolutionary selection pressures that underlie them, are summarized here.

In order to function as a specific regulator of gene expression in transcription, a regulatory protein must be able to recognize and to bind preferentially to certain DNA sequences against the background of a large number of competing sites of similar sequence that are also present in the genome. To this end the protein must have been selected during evolution to present a proper binding surface that can interact much more favorably with some DNA sequences than with others. The specific DNA recognition sequences must have been selected in a parallel evolutionary process to provide physiologically appropriate binding sites for the protein. The specific recognition (*binding selection*) of the DNA sequences by the protein is based on the physiochemical interactions between them. The *evolutionary selection* of these recognition properties is based on the requirement for a functionally adequate level of binding. Thus binding selection and evolutionary selection are strongly interrelated and this must be manifest in the statistics of base-pair usage that characterizes the set of specific recognition sites used by a particular regulatory protein.

Recognition

In what follows we shall discuss recognition mostly in terms of equilibrium binding affinity, although specific activity may be a more appropriate parameter, for example, for RNA polymerase binding to promoters, where functional recognition can also be based on activation steps such as the 'opening' of a specific region of the DNA double helix. When activation plays a role one must consider not just the binding free energies, but a combination of binding and activation free energies.

Primary sequence recognition: The recognition elements of DNA consist of the four different base pairs: A·T, T·A, C·G and G·C. In a protein binding site these base pairs are arranged in specified order so that they present a binding surface that is complementary to that of the protein. The individual base pairs expose patterns of hydrogen-bond donors and acceptors in the grooves of the DNA double helix to form a distinguishing matrix of interaction elements. Hydrogen-bond donors and acceptors are clearly the major recognition elements in DNA–protein interactions, since only these moieties can complement the hydrogen-bond acceptors and donors of the protein active site to provide a recognition matrix schematically is through a set of 'stick figure' representations of ation required[1–4]. A convenient way of representing the potential interaction matrix schematically is through a set of 'stick figure' representations of sequences of DNA base pairs[5].

When a base pair is replaced in a recognition sequence, thereby changing the complementarity of potential hydrogen-bond formation with the protein surface, binding affinity can be greatly affected. The loss of a protein–nucleic acid hydrogen bond can cause a reduction in favorable binding free energy of up to 16.8 kJ (4 kcal) mol^{-1} (see, for example Ref. 6) if there are no compensatory effects, and perhaps by 2·1–4·2 kJ (0.5–1 kcal) mol^{-1} if the molecular complex can rearrange to allow some compensatory hydrogen-bond formation, either internally or with solvent (see Ref. 7 for discussion).

Secondary sequence recognition: Interactions other than hydrogen bonds can also contribute to binding and specificity. Thus, hydrophobic interactions (especially with the methyl group of thymine), the steric fit of DNA groove geometries, and the conformation and flexibility of the DNA helix, are some of the elements that can influence protein binding in a sequence-dependent way. However, they are not sufficiently limiting by themselves to define a specific recognition sequence for a protein. Thus these elements serve mostly as binding modulators within the context of a hydrogen-bond matrix, or to favor binding of small molecules (e.g. certain drugs, see Ref. 8) to particular *regions* of the DNA double helix, rather than to define specific binding sequences *per se*. Ionic interactions of basic amino acid residues, involving counter-ion displacement from the negatively charged phosphates of the DNA backbone of the double helix[9], provide a significant share of the favorable binding free energy of proteins to nucleic acids in specific complex formation, and virtually all the binding free energy in nonspecific binding interactions (see below).

Distortions in the DNA, e.g. changes in the groove geometry or local bending of the helix, can also occur and can thermodynamically 'improve' protein binding. Such distortions cost free energy and therefore reduce the gain in binding affinity that might otherwise be realized. Local helix conformation and flexibility are strongly dependent on DNA sequence and on the choice of consecutive base pairs. Thus base pairs that influence conformation can contribute to specificity without contributing specific interactions with the protein (see, for example Refs 10 and 11).

Different binding modes: Many DNA-binding proteins, when faced with a DNA sequence lacking sufficient complementarity, will simply 'relax' to a purely nonspecific binding mode that is based primarily on electrostatic inter-

O. G. Berg is at the Department of Molecular Biology, University of Uppsala Biomedical Center, Box 590, S-75124 Uppsala, Sweden. P. H. von Hippel is at the Institute of Molecular Biology and Department of Chemistry, University of Oregon, Eugene, OR 97403, USA.

actions and is therefore largely sequence independent (see Ref. 12). This nonspecific binding mode may play important roles, both thermodynamically as a 'buffer' to control the free protein concentration and thus the equilibrium selection of specific sites (e.g. Ref. 13), and kinetically to regulate the rate of location of the specific DNA target sites[14]. The protein can also have more than one *specific* binding mode[15]; thus it can change conformation or distort to adjust to different functional DNA binding sequences. As a consequence, the protein will be able to interact favorably with a larger and more diverse set of DNA sequences and, in effect, will lose overall specificity. It is important to stress that conformational lability (or distortion) of the protein can only *reduce* specificity, while similar conformational lability in the DNA can be sequence dependent and therefore can *increase* specificity (see below).

Functional requirements

Occupancy of specific sites: The functional usefulness of a specific DNA recognition sequence is determined by the extent of protein binding under physiological conditions. This, in turn, is determined not only by the specific affinity, but also by the concentration of free protein available and by the concentration and affinity of alternative DNA sites to which the protein can bind in competition with the biologically relevant target sites.

Pseudosites: Since the genome is so large (some 10^6–10^7 base pairs in a bacterium like *E. coli*), one expects that a large number of sequences will occur by chance that are identical with, or very similar to, the biologically active sites. The probability of occurrence of such 'pseudosites' is determined primarily by the size of the DNA target site; i.e. by the number of base pairs with which the protein interacts specifically[4,12].

The binding affinity of a regulatory protein for a pseudosite is decreased relative to that for a specific site as a function of the number of 'wrong' (noncomplementary with the protein hydrogen-bonding matrix) base pairs that are included in the pseudosite. At the same time the expected number of random occurrences of a pseudosite with a certain affinity increases extremely rapidly with the number of noncomplementary base pairs allowed. Thus to keep binding of protein to pseudosites at reasonable levels it is important that the number of such sites be kept sufficiently low. This can be achieved by *overspecification* of the specific site; in effect by making the size of the specific sequence that defines the site larger than is strictly necessary to provide statistical uniqueness within the genome[4,12]. As a possible consequence not all potentially favorable interactions can be allowed to contribute to avoid making the specific binding *too* strong, and thus making dissociation times unreasonably long compared to overall metabolic and cell cycle rates. This imposes constraints on the types and strengths of interactions used in binding. Another way of keeping the competitive influence of pseudosites low is to make regulatory regions more accessible for regulatory protein binding than are other regions, perhaps by structural arrangements at the chromosome level (e.g. placement of nucleosomes).

To permit the correct physiological response to the same protein at functionally distinct recognition sites (e.g. at different promoters), it may be important that the extent of binding be set at well-defined (and different) levels for the various sites. This introduces variability in the specific sequences and can also impose constraints on the strengths of the DNA–protein interactions used.

Functional specificity: In essence, competitive binding at nonspecific and pseudospecific sites can be compensated by the cell through mass action by making a larger investment in protein. The functional specificity within the cell can thus be defined by the amount of protein that must be 'wasted' by nonproductive binding at 'wrong' sites in order to achieve the required physiological binding levels at specific sites. In some sense, therefore, one can expect that the level of functional specificity is determined by the physiological 'cost' of the protein invested. Obviously, because there would be no selective advantage for specificity without it, this competition with nonspecific sites and pseudosites must play an important role in the evolutionary selection of the recognition sites and of the proteins that recognize them.

Base-pair statistics of regulatory binding sites

For most proteins that regulate gene expression there exists a family of different specific recognition sites in the genome that serve different functions; e.g. as operator or promoter sites for different genes. These sites differ somewhat in sequence, but generally retain some common base-pair pattern that permits their recognition by the binding protein (and by the researcher). The choice of base pairs within these sites is determined by the potential for protein–DNA interactions that these base pairs represent, via the functional requirements discussed above. Therefore the statistics of base-pair choice within identified recognition sequences can be expected to carry information both about individual base-pair interactions and about the overall functional constraints that the sequences have been selected to satisfy by evolution.

Consensus sequence: By examining the family of specific DNA sites to which a certain protein binds one can deduce a consensus sequence by identifying the most commonly occurring base pair at each position in the sites. Departures from the consensus sequence correspond mostly (but not always) to 'down' mutations; i.e. to lessened function (see, for examples Refs 16 and 17). A strong correlation between function (defined as *in vitro* promoter strength) and closeness in sequence homology with the consensus sequence has been demonstrated for the promoter sites of *E. coli*[16,17]. Starting from basic principles, and making some simple assumptions, we have been able to derive general relations quantitating this correlation between function and sequence homology for a number of protein–DNA interaction systems[18,19].

Selection theory: All members of a family of DNA sites will display binding affinity (activity) for the relevant regulatory protein within a range that is defined by the amount of free protein in the cell and the required binding (activity) levels. The statistics of base-pair usage within this family of sites will be determined by the average binding affinity of the functional sites and by the contributions to binding of the individual base pairs. This is in total analogy to the situation in statistical thermodynamics, where the statistics of energy level occupancy is determined by the individual energy levels and the average or total energy available to the system.

These results can be derived under the convenient (though strictly neither necessary nor completely true) assumptions that: (1) base pairs contribute independently and therefore additively to the total binding free

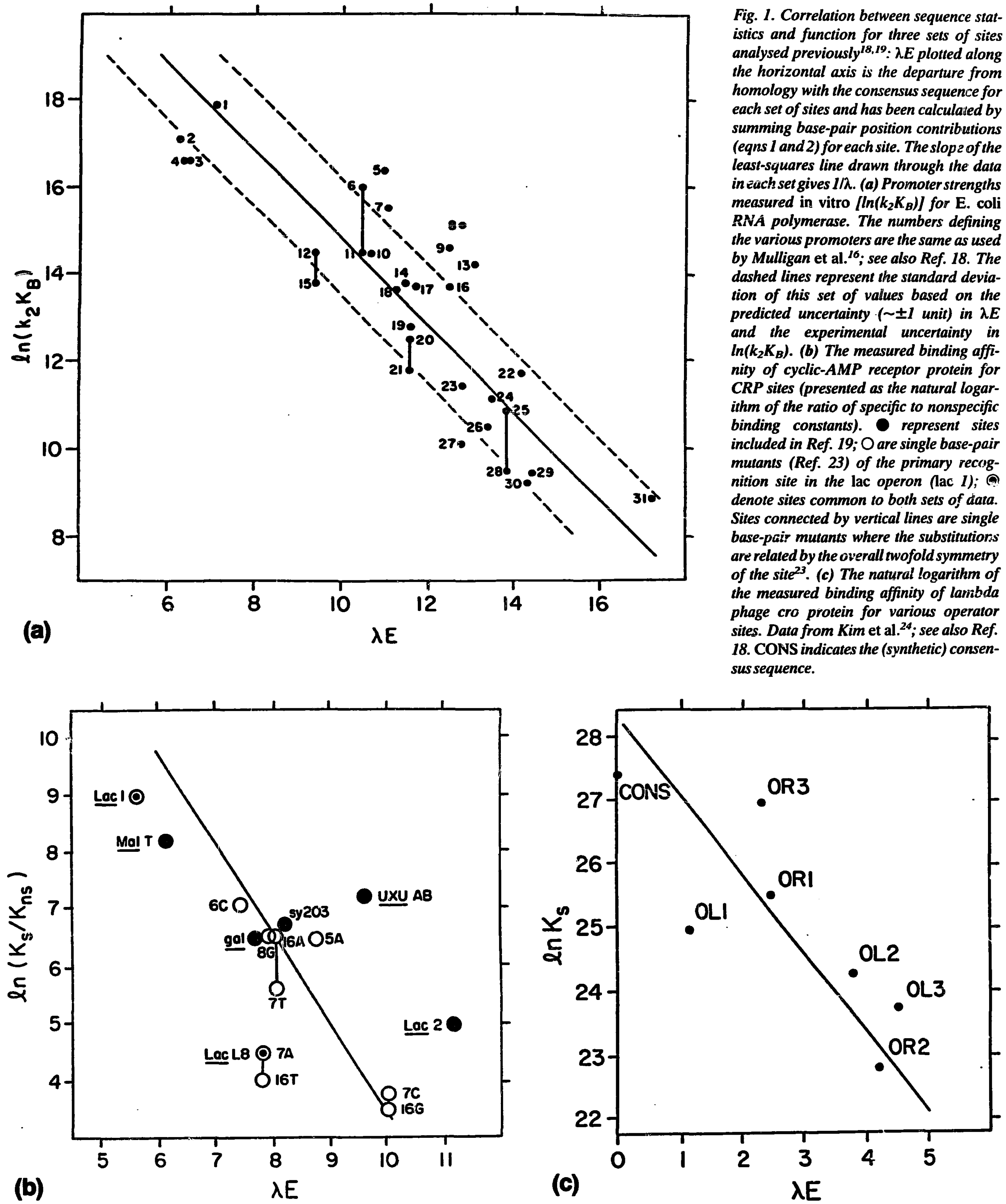

Fig. 1. Correlation between sequence statistics and function for three sets of sites analysed previously[18,19]: λE plotted along the horizontal axis is the departure from homology with the consensus sequence for each set of sites and has been calculated by summing base-pair position contributions (eqns 1 and 2) for each site. The slope of the least-squares line drawn through the data in each set gives 1/λ. (a) Promoter strengths measured in vitro *[ln(k_2K_B)] for* E. coli *RNA polymerase. The numbers defining the various promoters are the same as used by Mulligan* et al.*[16]; see also Ref. 18. The dashed lines represent the standard deviation of this set of values based on the predicted uncertainty (~±1 unit) in λE and the experimental uncertainty in ln(k_2K_B). (b) The measured binding affinity of cyclic-AMP receptor protein for CRP sites (presented as the natural logarithm of the ratio of specific to nonspecific binding constants). ● represent sites included in Ref. 19; ○ are single base-pair mutants (Ref. 23) of the primary recognition site in the* lac *operon (*lac *1); ◉ denote sites common to both sets of data. Sites connected by vertical lines are single base-pair mutants where the substitutions are related by the overall twofold symmetry of the site[23]. (c) The natural logarithm of the measured binding affinity of lambda phage cro protein for various operator sites. Data from Kim* et al.*[24]; see also Ref. 18.* CONS *indicates the (synthetic) consensus sequence.*

energy; and (2) all sites with the same binding affinity are equally likely to have been chosen as recognition sites during evolution. In this way it is possible to relate the frequency of occurrence of a certain base pair at a certain position in a site to a 'Boltzmann factor'. Conversely, the logarithm of the ratio of the frequencies of occurrence of different base pairs at a certain position in a set of sites will be proportional to the difference in binding free energy contributed by these respective base pairs to the overall binding affinity.

Thus if $n_{i,B}$ denotes the number of times base pair B (A·T, T·A, C·G or G·C) occurs at position *i* in the family of sites, one finds that the reduction ($\varepsilon_{i,B}$) in the free energy of binding (in units of *kT*) when base pair B replaces the consensus base pair at this position is (Eqn 1):

$$\lambda\varepsilon_{i,B} = \ln[PB(n_{i,0}+0.5)/P_0(n_{i,B}+0.5)]$$

where $n_{i,0}$, is the number of times that the consensus base pair occurs in position *i*; (the sign has been chosen such that $\varepsilon_{i,B}$ is a positive quantity when the binding is weaker). P_B/P_0 is the ratio of the overall genomic frequencies of base pair B and the particular consensus base pair. The terms 0.5 are 'small-sample' corrections that are particularly important when $n_{i,B}$ is small or even zero (see Ref. 19). The proportionality constant (λ) is a selection parameter that appears analogously to $1/kT$ in statistical thermodynamics; it can be shown to be related to the rate of change of the density of random pseudosites with decreasing affinity and possibly to the evolutionary selection pressure, as discussed further below.

Statistics–function correlations: Consequently, using the statistics of base-pair occurrence within a family of identified binding sites, it is possible to calculate the contributions to the binding free energy of any base pair at any position within the sites. By adding these contributions:

$$E = \sum_i (2)\varepsilon_{i,B} \qquad (2)$$

one can estimate E, the relative binding free energy for any base-pair *sequence* (apart from the unknown selection parameter, λ). By comparing the real (measured) and the estimated binding free energies for some sites, the absolute binding free energies and the selection parameter can be determined. It should be pointed out that E is the reduction in binding free energy, while the combination λE is a purely statistical measure expressing the departure of the sequence under consideration from the consensus sequence.

We have applied our selection theory to various families of recognition sites and find reasonable agreement. In general, known families of sites have only limited numbers of members, and the above procedure then carries a very large statistical 'small-sample' uncertainty. Nevertheless, we are encouraged by the fact that the agreement found is mostly within this uncertainty (see Fig. 1). In this figure we summarize the correlation of the *measured* function (or binding affinity) with λE calculated from base-pair statistics correlations for *E. coli* promoter sites, for cyclic-AMP regulatory protein (CRP) activation sites, and for lambda phage cro protein binding sites. Clearly, within the statistical uncertainty, the correlation between measured binding affinities (or activities) and those predicted from the selection theory (λE) are good for all these systems.

General properties: While estimates for individual sequences carry a large uncertainty, large deviations from the expected values can be an indication that specificity is strongly influenced by factors other than primary sequence. Thus the main CRP site in the *cat* operon represents an example of a site for which the primary sequence predicts no binding, although this site is known to be functional *in vivo* (see Ref. 19). (Perhaps such sites have evolved separately from a different precursor sequence; see, for example Ref. 20 for approaches to the identification of such sequences by random DNA synthesis techniques.)

Perhaps the most useful information that can be derived by this approach refers to the properties of whole families of sites. From the selection theory it has been possible to relate the expected extent of pseudosite competition directly to the sequence information carried in the specific sites (defined and calculated in Ref. 21 for various families of operator sites). As a consequence, one can estimate the functional specificity and the distribution of the protein over various sites *in vivo*. It appears that the functional specificity is low in many cases; i.e. a large fraction of the protein is 'wasted' by non-productive binding at nonspecific sites and pseudosites[19].

Many other aspects of specific DNA–protein interactions can be studied using this approach. For example, the effects of sequence asymmetry within the sites can be examined. Although the cyclic-AMP receptor protein interacts in the same way with particular base pairs at symmetrical positions in the two halves of its binding site[22], it seems that the functional sites have been selected to have weaker homology and therefore presumably weaker interactions with one half of the site[19]. No such asymmetry is found, for instance, in the recognition sites for the *lex*A protein[25].

One can use large samples (large families of known specific DNA binding sites) to examine correlations in the choice of neighboring base pairs within the binding sequences. Of the families of binding sites discussed here only the promoter sites, with over 100 different sequences identified, are sufficiently numerous to show significant correlations. No correlations (above the small-sample variations) can be detected at most positions *within* promoter sites, thus partially justifying the assumption of independent contributions from individual base pairs. In a few cases very strong nearest neighbor correlations do appear in the promoters; the strongest of these is a TG doublet just upstream of the −10 region, where the individual base pairs are only weakly conserved. Based on this strong correlation, we suggested[18] that this doublet might be an important recognition element. It has recently been experimentally identified as such[26,27]. This second strongest correlation constitutes a signal for stringent control of ribosomal promoters and is probably not directly related to the primary function of the RNA polymerase (see Ref. 18). Thus, apart from the individual base-pair preferences used to identify consensus sequences, base-pair correlations can also carry important information about DNA recognition elements.

What is λ? The selection parameter λ serves as a coupling factor between a purely statistical measure of deviation from base-pair randomness of regulatory sites in DNA and the functional (or binding) properties of recognition sequences. For a particular family of recognition sequences it can be determined as the inverse of the slope of a correlation line like those shown in Fig. 1.

λ is also related to the density of random pseudosites; in fact it can be shown that the number of random pseudosites with an affinity comparable to the specific sites increases proportionally to exp (λE) with a decrease of binding free energy, E (where E is expressed in units of kT and counted as positive for weaker binding). In the statistical thermodynamic analogy, λ is referred to as a 'generalized force'. For the specific sites it represents the strength of the tendency with which random mutations drive specific sites towards weaker specificity. Thus λ is the 'randomization pressure' due to errors in DNA replication that must be counteracted and balanced by the 'evolutionary selection pressure' that continuously removes organisms in which random mutations have driven recognition sites to nonfunctional levels of specificity. We conclude that λ may be considered to represent the force that is required to 'push' a set of random sequences to provide the required specific binding properties.

In the families of recognition sites described above we have found some significantly different values for λ. The relatively low value of λ for the recognition sites for the cyclic-AMP receptor protein may represent a real difference in the selection pressure on activation sites compared (for example) to promoter sites. This difference can be rationalized, since a change in affinity of an activation site may have only a minor influence on its biological function if this parameter depends on equilibrium binding occupancy. Thus if the level of occupancy of the activation site is (say) 0.9, a tenfold decrease in affinity would decrease occupancy to 0.47; i.e. by less than twofold. This is in contrast to the situation with promoter and operator sites, where a change in affinity influences biological activity in direct proportion[19].

It may also be significant that the selection parameter λ is close to 1.0 in most cases studied. This would be the expected result if these binding sites had been selected in evolution with the same bias as in a direct binding experiment. Equivalently, it represents a situation where the competitive binding to pseudosites has a maximum in the same region of affinity as the specific sites.

Conclusions

We conclude that it is possible to formulate a coherent picture of protein–DNA specificity, including both thermodynamic (binding) and evolutionary selection. While the procedures used to identify individual recognition properties from the statistics of base-pair deviations from consensus sequences does lead to large uncertainties, the formalism we have developed provides a useful framework for the discussion and evaluation of specificity. In particular it provides a direct estimate for the competitive influence of pseudosites. It can also suggest what binding properties are important for function, predict the energetic consequences of site specific mutagenesis of base pairs within the DNA target site and may even yield a quantitative estimate of the evolutionary selection pressure on the particular control circuits in which the recognition sites are involved. We look forward to further tests and application of these approaches to biological specificity and its evolutionary development.

Acknowledgements

This work was supported in part by USPHS research grants GM–15792 and GM–29158 (to PHvH). Salary support for OGB was provided by the Swedish Natural Science Research Council.

References

1 Yarus, M. (1969) *Annu. Rev. Biochem.* 38, 841–880

2 von Hippel, P. H. and McGhee, J. D. (1972) *Annu. Rev. Biochem.* 41, 231–300

3 Seeman, N. C., Rosenberg, J. M. and Rich, A. (1976) *Proc. Natl Acad. Sci. USA* 73, 804–808

4 von Hippel, P. H. (1979) in *Biological Regulation and Development* (Goldberger, R. F., ed.), pp. 279–347, Plenum Press

5 Woodbury, C. P., Jr and von Hippel, P. H. (1981) in *The Restriction Enzymes,* (Chirikjian, J., ed.), pp. 181–207, Elsevier

6 Tronrud, D. E., Holden, H. M. and Matthews, B. W. (1987) *Science* 235, 571–574

7 Fersht, A. R. (1987) *Trends Biochem. Sci.* 12, 301–304

8 Kopka, M. L., Yoon, C., Goodsell, D., Pjura, P. and Dickerson, R. E. (1985) *Proc. Natl Acad. Sci. USA* 82, 1376–1380

9 Record, M. T., Jr, Lohman, T. M. and deHaseth, P.L. (1976) *J. Mol. Biol.* 107, 145–156

10 Liu-Johnson, H-N., Gartenberg, M. R. and Crothers, D. M. (1986) *Cell* 47, 995–1005

11 Anderson, J. E., Ptashne, M. and Harrison, S. C. (1987) *Nature* 326, 846–852

12 von Hippel, P. H. and Berg, O. G. (1986) *Proc. Natl Acad. Sci. USA* 83, 1608–1612

13 von Hippel, P. H., Revzin, A., Gross, C. A. and Wang, A. C. (1974) *Proc. Natl Acad. Sci. USA* 71, 4808–4812

14 Berg, O. G., Winter, R. B. and von Hippel, P. H. (1982) *Trends Biochem. Sci.* 7, 52–55

15 Mossing, M. C. and Record, M. T., Jr (1985) *J. Mol. Biol.* 186, 295–305

16 Mulligan, M. E., Hawley, D. K., Entriken, R. and McClure, W. R. (1984) *Nucleic Acids Res.* 12, 789–800

17 Mulligan, M. E. and McClure, W. R. (1986) *Nucleic Acids Res.* 14, 109–126

18 Berg, O. G. and von Hippel, P. H. (1987) *J. Mol. Biol.* 193, 723–750

19 Berg, O. G. and von Hippel, P. H. *J. Mol. Biol.* (in press)

20 Horwitz, M. S. Z. and Loeb, L. A. (1986) *Proc. Natl Acad. Sci. USA,* 83, 7405–7409

21 Schneider, T. D., Stormo, G. D., Gold, L. and Ehrenfeucht, A. (1986) *J. Mol. Biol.* 188, 415–431

22 Ohlendorf, D. H., Anderson, W. F., Fisher, R. G., Takeda, Y. and Matthews, B. W. (1982) *Nature* 298, 718–723

23 Ebright, R. H., Kolb, A., Buc, H., Kunkel, T. H., Krakow, J. S. and Beckwith, J. (1987) *Proc. Natl Acad. Sci. USA* 84, 6083–6087

24 Kim, J. G., Takeda, Y., Matthews, B. W., and Anderson, W. F. (1987) *J. Mol. Biol.* 196, 149–158

25 Berg, O. G. *Nucleic Acids Res.* (in press)

26 Keilty, S. and Rosenberg, M. (1987) *J. Biol. Chem.* 262, 6389–6395

26 Ponnambalam, S., Chan, B. and Busby, S. (1988) *Mol. Microbiol.* 2, 165–172