Specificity, free energy and information content in protein–DNA interactions

Gary D. Stormo and Dana S. Fields

Site-specific DNA–protein interactions can be studied using experimental and computational methods. Experimental approaches typically analyze a protein–DNA interaction by measuring the free energy of binding under a variety of conditions. Computational methods focus on alignments of known binding sites for a protein, and, from these alignments, make estimates of the binding energy. Understanding the relationship between these two perspectives, and finding ways to improve both, is a major challenge of modern molecular biology.

TRANSCRIPTIONAL REGULATION IS mediated, in part, by alteration of the promoter activity by proteins that bind to sites on DNA. Importantly, the sites that are recognized by any one DNA-binding protein are in general not a unique sequence (as one sees for some restriction enzymes). Rather, the sites of recognition are a family of similar sequences, and, naturally, non-specific binding sites (non-sites) for the protein are the collection of sequences that do not fall into the protein's family of recognition sequences. Since the vast majority of the genome comprises non-site DNA sequences, and since site-specific DNA-binding proteins still have a weak affinity for the non-site DNA, the protein must display a much higher binding affinity for its own site(s) than for non-site DNA in order for the regulatory system to work.

Within the cell, or nucleus for eukaryotic systems, the concentration of DNA is so high that the protein will be bound to DNA, site or non-site, essentially all of the time. It is the ability of the protein to distinguish its proper binding site(s) from the rest of the non-site DNA that is essential for the proper functioning of the regulatory system: this ability to distinguish site from non-site is called specificity. (An excellent presentation on general issues of specificity can be found in a paper by von Hippel[1].) Specificity in itself does not require, as one might have thought, that the protein display a large absolute affinity for its binding site(s), although having such a property does make the binding reaction easier to study *in vitro*.

G. D. Stormo and **D. S. Fields** are in the Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, CO 80309-0347, USA.
Email: Stormo@Colorado.Edu

Several DNA-binding motifs[2–5] have been studied in depth, in part using crystal structures of protein–DNA complexes. This work has been complemented by experiments that measure the affinity of a DNA–protein interaction as the equilibrium constant of a binding reaction[6–8]. Taken together, some general principles that govern DNA–protein interactions have emerged[9–11]. In particular, the search for a 'DNA recognition code', which sets out rules for amino acid–base pair interactions, has shown that the code is not like the deterministic genetic code. Rather, the rules that are emerging specify preferences for base pair–amino acid interaction and it is these preferences that govern specificity. In this paper, we do not discuss the recognition code *per se*, but we do present the notion of specificity in terms of equilibrium binding constants and information content.

Quantitative specificity

The simplest model of a DNA–protein binding reaction is given by the equation: $\mathbf{T}+X_i \leftrightarrow \mathbf{T}\bullet X_i$. Here **T** is free protein, X_i is any one particular site or non-site DNA, and $\mathbf{T}\bullet X_i$ is bound complex. A convenient way to quantify the specificity of such a protein is to normalize the binding constant:

$$K_{\mathrm{eq}} = \frac{[\mathbf{T} \bullet \mathrm{X}_i]}{[\mathbf{T}][\mathrm{X}_i]}$$

to some user-defined reference value. One typical reference is the K_{eq} of the 'preferred' sequence: that is, the highest value of K_{eq} over all the X_i. In this way, the reference sequence has a specificity of 1 by definition and the specificity of all the other sequences falls between 1 and 0, with the non-sites being very close to 0.

We also normalize binding constants to an average of the set of K_{eq} over all the X_i, yielding 'specific binding constants', K_{S}.* This is useful because it presents specificity values with respect to the binding behavior of the genome as a whole. Using specific binding constants, preferred sequences (i.e. bona fide binding sites) will have a very large value of specificity, perhaps 10^6. Most sequences (e.g. non-sites), however, will have values less than 1 because the average is composed with a vast excess of non-site terms. In *Escherichia coli*, with a genome of about 5×10^6 bp, a protein with a specificity of 10^6 for a particular sequence would be bound to that site only about 20% of the time. Still, it would only take about 20 copies of the protein in the cell to maintain occupancy of the site at about 99% (Ref. 12).

Our definition of K_{S} is very convenient for a thermodynamic analysis of regulation, because the probability that a protein will be bound at a particular site is equal to the K_{S} for the site divided by the partition function, Z, which is the sum of the K_{S}s for all possible sites in the genome. Since K_{S} is defined to have an average value of 1 for all sites in the genome, the partition function is just the size of the genome, Γ. Thus, the probability that a protein is bound to a particular site, X_i, when it has the possibility of binding anywhere in the genome, is just $K_{\mathrm{S}}(X_i)/\Gamma$. (Note that this is only the probability of a single protein binding at a site, and that calculation of the probability that the site is occupied also depends on the number of proteins in the cell[12].)

*'All the X_i' may be defined differently, depending on one's purpose. For example, it may be useful to consider all of the possible binding sequences for a protein, even those that do not occur in a particular genome. If the protein recognizes sites that are L bases long, there are 4^L possible sequences. A typical prokaryotic regulatory protein binds to sites about 20 bases long, so there are $4^{20} \approx 10^{12}$ such sequences, many more than the size of even the largest genomes. Measuring specificity with respect to all possible sequences provides a convenient means of comparing the specificity of different proteins from different systems because there is an absolute scale of comparison. However, for understanding how a regulatory system works *in vivo*, it is more appropriate for 'all the X_i' to refer to the collection of sequences that occur in the genome of the organism. That is the approach taken in this paper.

0968–0004/98/$19.00 PII: S0968-0004(98)01187-6

The K_S for a site could be determined directly by experimentally measuring the relative affinity of the protein for the site compared with the whole genome. But for many purposes one would like to know how the affinity of the protein varies for different sequences – ideally, the affinity for all possible sites in the genome. However, this is impractical even for proteins that recognize relatively short sequences. But it is feasible to measure the binding affinity for a collection of sequences that are all similar to the known binding site(s) (e.g. mutants of the wild-type or consensus sequence), and then to extrapolate these data to all the other sequences that were not measured directly. For this extrapolation process to be of use, one assumes that the positions in the binding site contribute independently to the binding energy. This means, for example, that the change in the binding energy of a double mutant of the binding site is just the sum of the changes from each of the two respective single mutants of the same site. Under this independence assumption, a table of the K_S values (or, equivalently, $\Delta G_S = -RT \ln K_S$) for all the single mutants of a binding site allows one to estimate the affinity for any site or non-site sequence. This independence assumption is not likely to be exactly true in general; however, it seems to be a reasonable approximation in many cases[13–16]. One way to monitor the assumption's validity is to perform an experiment[12,17] in which the protein is allowed to select its own binding site(s) from a randomized pool of DNA using SELEX (Ref. 18; Box 1). It can then be readily determined from the recovered sequences whether or not there are correlations in the frequencies of bases at positions in the sites. Under the independence assumption there should be no correlations.

The most likely non-independent interactions will be between adjacent positions in the binding site. These near-neighbor effects may occur because the amino acid side-chains and the base pairs, perhaps involving water or other solvent molecules, can bind in a 'network' of interactions such that the energy of interaction with one base pair depends on the neighboring base pairs. Alternatively, the local structure of the DNA itself depends on the local sequence, and this may influence the binding in a way that is not additive. In cases where the additivity of single positions is not valid, extension of the approach to the dinucleotide or trinucleotide level may suffice. While this increases the complexity of the problem, it is still a great reduction of work when compared with the examination of all possible sequences. If the site recognized is L bases long, analysis of all single base changes requires the study of $3L + 1$ sites. Using adjacent dinucleotides increases that number to $15(L - 1) + 1$, which is still much smaller than all possible sites, 4^L.

Experimental measurements

In 1989 Sarai and co-workers published two papers examining the change in affinity for all possible single base changes away from the consensus *lambda* operator, for both of the proteins that bind to those operators, *cro* and *repressor*[15,16]. They also showed that the affinities of multiple mutants were reasonably well predicted based on the assumption of additivity. More recently, several proteins have been studied by measurement of the binding affinity to variants of the binding site, and also by selection of preferred sites from random sequences (see, for example, Refs 13, 14, 17, 29, 20). These results provide additional information about the specificity of the protein, and can be used in conjunction with the X-ray structures of the DNA–protein complexes to help understand the rules of recognition.

Box 1. SELEX

SELEX stands for the Systematic Evolution of Ligands by EXponential enrichment [see Gold, L. (1995) *J. Biol. Chem.* 270, 13581–13584]. In this cyclic process, a starting group of DNA or RNA sequences is made with a variable region in the center, surrounded by constant regions used for amplification by polymerase chain reaction (PCR). The sequence collection is then passed through a procedure whereby some subset group is purified away from the initial group, i.e. a subset is 'selected'. The selected subset is then amplified by PCR and that can again be subject to selection. This cyclic procedure can be repeated many times, until the purified subset meets some user-defined criteria. In our experiments with Mnt, the starting DNA was a sample of synthetic DNA which included a completely randomized region of 20 base pairs. That is, it included all possible 20-long sequences (about 10^{12} different sequences). The DNA was incubated with Mnt under DNA-binding conditions. The DNA sequences which bound were separated in a 'gel shift' from those that were not bound. The bound fraction was amplified by PCR and the selection step repeated. We did this for a total of nine rounds before cloning and sequencing the selected products; 62 sequences were obtained that were all very similar to the wild-type operator, but displayed some variability, which allowed us to determine the relative importance of each position, and the preference for each base at most positions.

One standard method of measuring binding affinities uses a 'gel shift' experiment. Recall that:

$$K(X_i) = \frac{[\mathbf{T} \bullet X_i]}{[\mathbf{T}][X_i]}$$

The ratio of bound $[\mathbf{T}\bullet X_i]$ to unbound $[X_i]$ DNA can be determined from the gel shift experiment, among other possible approaches. The concentration of free (and active) protein, $[\mathbf{T}]$, is usually not known and is in fact somewhat difficult to measure. Thus, typically, the ratio of bound to unbound DNA is determined at several different protein concentrations and then curve-fitting is employed to give a best estimate of $K(X_i)$. Multiple determinations of the same constant indicate that this approach can give values that are accurate to within a factor of about 2.

Focusing on the binding of the homotetrameric Mnt protein of *Salmonella* phage P22 to variants of its symmetric 17-base operator[19], we recently developed a new way to study how changes in sequence affect binding affinity that is both more rapid and more accurate than previous methods[12,21] (a similar approach was developed independently by Luo *et al.*[20]). Our method permits the simultaneous measurement of the relative binding constants of multiple variants of the binding site. Central to this method is the placement of multiple potential sites in the same binding reaction to force a competition between the different binding sites for the same pool of protein. Suppose then that we have two variants of a binding site, X_1 and X_2, and that we wish to know which is a better binder and by how much. If both DNAs are in the same test tube, hence competing for the same pool of protein, we can write the relative binding constant as:

$$\begin{aligned} \frac{K(X_1)}{K(X_2)} &= \frac{[X_1 \bullet \mathbf{T}]}{[X_1][\mathbf{T}]} \frac{[\mathbf{T}][X_2]}{[X_2 \bullet \mathbf{T}]} \\ &= \frac{[X_1 \bullet \mathbf{T}]}{[X_2 \bullet \mathbf{T}]} \frac{[X_2]}{[X_1]} \end{aligned} \tag{1}$$

Note that the free protein concentration has canceled out and that the binding reaction only needs to be performed once under a protein concentration that is chosen so that none of the relevant DNA concentrations, both free and in complexes, is zero (or undetectable). Furthermore, as shown in the equation, the relative binding constants (and trivially from this the K_S) can be determined by ascertaining the ratios of the two DNAs in both the bound and unbound samples. In addition, while the equation is only shown for two competing DNAs, any mixture of multiple DNAs can be analyzed simultaneously, provided only that each of them is detectable in both the bound and unbound fraction. In our study of Mnt, the 'multiple potential sites' are actually a single synthetic oligo-DNA sample of the wild-type *mnt* operator bearing one or two randomized positions. (To reduce the effects of non-independence in the operator, mutant operators never bear adjacent changes.)

A binding reaction of Mnt protein with this mixed oligo sample is performed and the bound and unbound fractions are separated with a gel shift. The most difficult part of the experiment is to determine accurately the ratio of each variant of the operator in the two fractions. Our first experiments of this type were designed so that the mixed oligo sample also contained different restriction enzyme sites[22]. In this way, we could determine the ratios of the variant operators in both the bound and unbound fractions with restriction digestions. However, this approach is too limited to be useful in general. Therefore we took the tools of dideoxy-DNA sequencing and used them to examine directly the ratios of the variant operators in both fractions. This 'quantitative sequencing' method[12,21] is powerful because, for a doubly randomized mutant of the operator, it can return 15 different relative binding constants (using the assumption of independence mentioned earlier) from one experiment. Moreover, the measured values display a very good precision, and repeated experiments with Mnt have shown that the values obtained are usually consistent to within about 20%, or 1.2-fold.

Table I shows our best estimates of K_S and ΔG_S of the Mnt protein for all possible single base changes from the wild-type *mnt* operator determined mostly by the 'quantitative sequencing' method[12,21]. (The wild-type sequence of the *mnt* operator is shown in Fig. 1b.) The wild-type base is the one with the highest affinity at every position except 19, where a C displayed an increased binding by about 1.3-fold above the wild-type T. The overall range of changes in binding is quite dramatic, from a 1.3-fold decrease for an A to C change at position 16, to about a 70-fold decrease for a G to C change at position 14.

As noted earlier, given a table of ΔG_S values at each position, as in Table I, and making the assumption of additivity, we can calculate a predicted ΔG_S for any sequence. We use the notation $\vec{\mathbf{G}}\bullet X_i$ for the calculation, where $\vec{\mathbf{G}}$ is the matrix of the ΔG_S values for each base at each position, and the sequence of X_i indicates the base to use at each position in the sum. Figure 1 provides an example using the wild-type sequence. In this case, the ΔG_S is -8.2 kcal mol^{-1} for the full site (the central position 11 only contributes once, but all other bold numbers are used twice), which corresponds to a K_S of about 6×10^5. This means that about 30 Mnt repressors are required within the cell to maintain about 99% occupancy of the operator[12].

Table I. Specific binding constants for the Mnt protein[a]

Position	Parameter[b]	*b* = A	*b* = C	*b* = G	*b* = T
11	$K_S(b)$	0.66	1.30	1.30	0.66
	$\Delta G_S(b)$	0.26	−0.16	−0.16	0.26
12	$K_S(b)$	0.55	0.26	3.00	0.15
	$\Delta G_S(b)$	0.37	0.87	−0.68	1.20
13	$K_S(b)$	0.23	0.39	0.56	2.80
	$\Delta G_S(b)$	0.91	0.58	0.36	−0.63
14	$K_S(b)$	0.60	0.045	3.00	0.33
	$\Delta G_S(b)$	0.31	1.90	−0.68	0.68
15	$K_S(b)$	0.91	0.30	2.50	0.25
	$\Delta G_S(b)$	0.058	0.74	−0.56	0.85
16	$K_S(b)$	1.60	1.20	0.48	0.81
	$\Delta G_S(b)$	−0.29	−0.11	0.45	0.13
17	$K_S(b)$	0.41	3.20	0.13	0.30
	$\Delta G_S(b)$	0.55	−0.72	1.30	0.74
18	$K_S(b)$	0.80	1.70	0.61	0.84
	$\Delta G_S(b)$	0.14	−0.33	0.30	0.11
19	$K_S(b)$	0.53	1.70	0.48	1.30
	$\Delta G_S(b)$	0.39	−0.33	0.45	−0.16

[a] The *mnt* operator is symmetric about position 11, so positions 12–19 serve to describe positions 3–10. Positions 1, 2, 20 and 21 are not included in the operator for this paper.

[b] Units for $\Delta G_S(b)$ are kcal mol^{-1}.

In this same study, we also selected Mnt binding sites from random DNA using the SELEX procedure mentioned above. The sequences so selected confirmed the relative importance of the different positions in the site, and also the relative preference of the different bases at each position. The SELEX data did reveal a statistically significant correlation at two positions, 16 and 17, suggesting that non-independent interactions occur there. Figure 2a shows the number of occurrences from the SELEX experiments for each base at positions 16 and 17. Note that the order of frequencies for each base corresponds closely to that expected from their specific binding constants in Table I. However, the occurrence of dinucleotide pairs at those positions deviates significantly from independence, as shown in Fig. 2b. While AC is the preferred dinucleotide, as expected from the mononucleotide preferences, the occurrences of several other pairs are quite unexpected. For example, when position 16 is not an A, position 17 is still preferentially a C, but only by a ratio of 24:7, compared to 93:0 when position 17 is an A. Even more surprising, when position 17 is not a C, position 16 is never an A, even though that would still be preferred if the positions were independent. It is clear that the very strong preference for C at position 17, as reported in Table I, depends on position 16 being an A, and that the slight preference for A at position 16 depends on a C at position 17. The exact magnitude of the non-independent contributions to the binding energy needs to be measured, and experiments are now under way to do this. We predict that those contributions will be smaller than the independent contributions because the preferences at each position are consistent with independence. However, the non-independence is certainly significant and needs to be taken into account for an accurate description of the total regulatory system.

Information and specificity

The large-scale genome sequencing projects currently under way, together with experiments to monitor the expression of all the genes under a variety of conditions, allow the use of pattern recognition methods to identify regulatory sites without doing binding experiments at all[23]. Given a collection of regulatory sites, one would like to obtain an estimate of the specificity of the DNA-binding protein directly from that sample of sites. Berg and von Hippel[24,25] first expressed the relationship between the statistics of example sites and the estimated binding energy. In what follows, we provide a somewhat different approach to the same problem, and show how the information in the aligned sites is used to estimate ΔG_S.

Suppose that we are given an aligned set of sequences, S_i, where we know that each S_i is a regulatory site for our DNA-binding protein. The set S_i is a very small subset

(a)

	11	12	13	14	15	16	17	18	19
	–	–	–	–	–	–	–	–	–
A	0.26	0.37	0.91	0.31	0.058	**–0.29**	0.55	0.14	0.39
C	–0.16	0.87	0.58	1.90	0.74	–0.11	**–0.72**	**–0.33**	–0.33
G	**–0.16**	**–0.68**	0.36	**–0.68**	**–0.56**	0.45	1.30	0.30	0.45
T	0.26	1.20	**–0.63**	0.68	0.85	0.13	0.74	0.11	**–0.16**

(b)

3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
a	*g*	*g*	*t*	*c*	*c*	*a*	*c*	*g*	*g*	*t*	*g*	*g*	*a*	*c*	*c*	*t*

Figure 1

(a) The collection of ΔG_Ss for the Mnt protein in Table I is presented as a matrix, $\vec{\mathbf{G}}$. Since the entire matrix is symmetric about column 11 only half is shown here. **(b)** A typical X_i: the wild-type *mnt* operator. In general, $\vec{\mathbf{G}} \bullet X_i$ picks out the ΔG_S in each column of $\vec{\mathbf{G}}$ that matches the base of the corresponding position in X_i, followed with a summation of these ΔG_Ss. That is, working with the half-site information, the sum over the bold-face entries in **(a)** is performed.

of the set we have been denoting as X_i, but, because they are regulatory sites, we can assume that each site in S_i has a high probability of being bound by the protein. The goal is to determine a matrix $\vec{\mathbf{W}}$, based only on the sample of known sites, which is a good estimate for the true $\vec{\mathbf{G}}$ of the protein. Two simplifying assumptions make this a straightforward process.

(a)

	Position 16	Position 17
	–	–
A	93	3
C	19	117
G	3	3
T	9	1
–	–	–
Wild type	93	117
Non-wild type	31	7

(b)

	*A*17	*C*17	*G*17	*T*17
	–	–	–	–
*A*16	0	93	0	0
*C*16	3	14	2	0
*G*16	0	2	1	0
*T*16	0	8	0	1

Figure 2

(a) The number of occurrences of each base at positions 16 and 17 in the *mnt* binding sites generated in the SELEX experiments is shown in the first four rows. The last two rows give the same information, with the four bases categorized as wild-type (A for 16, C for 17) and non-wild-type. **(b)** The number of occurrences of each dinucleotide combination at positions 16 and 17 in the *mnt* binding sites generated in the SELEX experiments.

The first assumption is one we have already been using, namely that the energy contributions are additive across the positions of the site. This allows us to represent the specificity by a mononucleotide matrix. This also means that we can summarize the information in the aligned binding sites by a simple table of the frequencies of each base at each position, $f(b, j)$. If we know $\vec{\mathbf{G}}$, then the average ΔG_S for all of the sites is just $\Sigma_{b,j} f(b,j)\Delta G_s(b,j)$ (or $\vec{\mathbf{f}} \bullet \vec{\mathbf{G}}$).

The second assumption is that the genome as a whole can be approximated as a random sequence. Clearly genomes are not random sequences, but the real issue here is whether the composition of the binding-site-sized non-sites is approximately random; that is, if the site size is ten bases, do all ten-base sequences occur with the frequency expected from the composition of the genome? While there are some clear biases in oligo composition, such as the CpG reduction in mammalian genomes, this assumption is usually a good approximation. (This assumption simplifies the estimate of the partition function, but it can be removed from the analysis with a concomitant increase in the complexity of the problem.)

Given any table of relative (or specific) binding constants $K(b, j)$, where b is each base and j the position in the site (see Table I), then the assumptions of independence and a random genome mean that the average binding constant, over all sequences, is just $\Pi_j \Sigma_b p(b)K(b,j)$, where $p(b)$ is the probability of each base in the genome. The partition function is that average times the size of the genome, Γ. Note that, if we use specific binding constants, K_S, as defined earlier, then we have already normalized the individual values at each position such that the average is 1 (consider Table I, using the fact that in *E. coli* each of the bases is at a probability of 0.25) and the partition function is just Γ.

Remember that our goal is to estimate $\vec{\mathbf{G}}$ by creating $\vec{\mathbf{W}}$, which is based only on a sample of the known sites. (Note: to keep with convention, the sign is switched between $\vec{\mathbf{W}}$ and $\vec{\mathbf{G}}$. $\vec{\mathbf{W}}$ is a 'weight matrix' based on the example sites and, by convention, sites that are preferred have higher scores, usually large positive numbers. Those same sites will have large negative values of ΔG_S.) The collection of known sites, which can be summarized in the table of frequencies for each base at each position $f(b,j)$, can be assumed to each have a high probability of binding to the protein. We can easily find the values for the matrix that maximizes the probability of binding to all of the sequences in the collection[26]. It is just:

$$\vec{\mathbf{W}}(b,j) = \log_2 \frac{f(b,j)}{p(b)}$$

[An additional important consideration is how to adjust for a limited sample of sites, in which case the $f(b,j)$ values may be biased[24,25].]

The *information content* of an aligned set of sites is defined as:

$$\begin{aligned} I_{\text{seq}} &= \sum_j \sum_b f(b,j) \log_2 \frac{f(b,j)}{p(b)} \\ &= \sum_j \sum_b f(b,j)\vec{\mathbf{W}}(b,j) = \vec{\mathbf{f}} \bullet \vec{\mathbf{W}} \end{aligned} \qquad (2)$$

These equations and the previous discussion emphasize that information content is an estimate of the average specific binding energy for the collection of known binding sites, using as the estimate of the binding function ($\vec{\mathbf{W}}$) that which maximizes the probability of binding to those sites. For an efficient regulatory system, the value of I_{seq} should be close to $\log_2\Gamma$, as is often observed, because otherwise too much of the protein will be bound to non-site DNA (Refs 12, 27). When using pattern recognition methods to discover regulatory sites in co-regulated genes, I_{seq} is a very useful objective function[23,26].

All of the analyses presented have assumed that the interaction under consideration is governed by equilibrium thermodynamics. A reaction that involves a complicated kinetic process may not have such a simple relationship between the activity of different sites and their degree of sequence conservation. For example, the activity of a promoter, as measured by the rate of RNA production, depends on several kinetic reactions following the initial binding of the polymerase, including the isomerization of the DNA to an open complex, the initiation of RNA polymerization and the release of the promoter DNA during the elongation phase. Given such a process it would not be surprising to find that the bases in the promoter-initiation region contributed in complicated, non-additive ways. However, *E. coli* promoters appear to be fairly well modeled by a simple relationship between sequence and activity. In fact, those were the sites first analyzed by Berg and von Hippel[24], and even earlier Mulligan *et al.*[28] had shown that a simple statistical measure of sequence conservation correlated well with promoter activity. In this case, the promoter activity was determined by an 'abortive initiation' assay which had been shown to be well approximated by the product of the equilibrium binding constant for the polymerase to the promoter, K_B, and the rate constant for isomerization of the DNA to the open complex, k_2 (Ref. 29). Perhaps it is because the activity is proportional to the simple product K_Bk_2 that the relationship to sequence is also simple, although it would not be too surprising if somewhat more complicated functions did somewhat better at modeling the activity. In other cases, such as *E. coli* ribosome binding sites, there are interactions between different rate constants that require more complicated models to represent accurately the relationship between sequence and activity[30].

Conclusions and future directions

Understanding specificity in DNA–protein interactions is essential for a thorough understanding of gene regulation. Much progress has been made in recent years through a combination of structure determination of DNA–protein complexes and methods for assessing differences in binding affinity for different sequences. We have presented a new approach for measuring specific binding interactions that is both rapid and accurate. More information is still needed about the validity of the additivity assumption in general, and about how to recognize cases when it is not valid. We also need to know whether, in those cases, it would be sufficient to use dinucleotide additivity models, or if even higher levels would be required. We have also shown that sets of example sites can be used to estimate ΔG_S for the protein recognizing those sites. We would like to be able to relax the simplifying assumptions without increasing the complexity of the problem too much, and we would like to be able to handle more effectively cases with only a few example sites. Improvements to the algorithms as well as better experimental methods can help reduce these limitations.

Finally, we need to delve more deeply into complex regulatory systems. Our approach up to now has been very reductionist, looking at individual components of the system and analyzing how they interact. But *in vivo* the system can be much more complicated. Often multiple proteins are required for proper regulation, and knowledge of how they interact with each other, as well as with the DNA, is essential to the total story. How different proteins compete for the same, or overlapping, sites and how chromatin structure influences the binding activity are important questions. These are active areas of research that we have not touched upon in this brief article. But it is clear that a true understanding of regulatory systems requires analysis at all these levels. Current large-scale sequencing projects, as well as new methods to monitor gene expression, promise to provide us with many examples of regulatory systems. Unraveling those, through combinations of molecular biology and mathematical algorithms, is a great challenge. It is certainly an exciting time to be working in this field.

References

1 von Hippel, P. H. (1979) in *Biological Regulation and Development*, Vol. 1 (Goldberger, R. F., ed.), pp. 279–347, Plenum
2 Pabo, C. O. and Sauer, R. T. (1992) *Annu. Rev. Biochem.* 61, 1053–1095
3 Phillips, S. E. V. (1994) *Annu. Rev. Biophys. Biomol. Struct.* 23, 671–701
4 Nelson, H. C. M. (1995) *Curr. Opin. Genet. Dev.* 5, 180–189
5 Hendrickson, W. and Blundell, T. L., eds (1996) *Curr. Opin. Struct. Biol.* 6(1)
6 Sauer, R. T., ed. (1991) *Methods Enzymol.* 208
7 Shadle, S. E., Allen, D. F., Guo, H., Pogozelski, W. K., Bashkin, J. and Tullius, T. D. (1997) *Nucleic Acids Res.* 25, 850–860
8 Ragnhildstveit, E., Fjose, A., Becker, P. B. and Quivy, J-P. (1997) *Nucleic Acids Res.* 25, 453–454
9 Choo, Y. and Klug, A. (1997) *Curr. Opin. Struct. Biol.* 7, 117–125
10 Mandel-Gutfreund, Y., Schueler, O. and Margalit, H. (1995) *J. Mol. Biol.* 253, 370–382
11 Suzuki, M. (1994) *Structure* 2, 317–326
12 Fields, D. S., He., Y., Al-Uzri, A. and Stormo, G. D. (1997) *J. Mol. Biol.* 271, 178–194
13 Lustig, B. and Jernigan, R. L. (1995) *Nucleic Acids Res.* 23, 4707–4711
14 Desjarlais, J. R. and Berg, J. M. (1994) *Proc. Natl. Acad. Sci. U. S. A.* 91, 11099–11103
15 Takeda, Y., Sarai, A. and Rivera, V. M. (1989) *Proc. Natl. Acad. Sci. U. S. A.* 86, 439–443
16 Sarai, A. and Takeda, Y. (1989) *Proc. Natl. Acad. Sci. U. S. A.* 86, 6513–6517
17 He, Y., Stockley, P. G. and Gold, L. (1996) *J. Mol. Biol.* 255, 55–66
18 Gold, L. (1995) *J. Biol. Chem.* 270, 13581–13584
19 Knight, K. L. and Sauer, R. T. (1992) *EMBO J.* 11, 215–223
20 Luo, B., Perry, D. J., Zhang, L., Kharat, I., Basic, M. and Fagan, J. B. (1997) *J. Mol. Biol.* 266, 479–492
21 Fields, D. S. and Stormo, G. D. (1994) *Anal. Biochem.* 219, 230–239
22 Stormo, G. D. and Yoshioka, M. (1991) *Proc. Natl. Acad. Sci. U. S. A.* 88, 5699–5703
23 Hertz, G. Z., Hartzell, G. W. and Stormo, G. D. (1990) *Comput. Appl. Biosci.* 6, 81–92
24 Berg, O. G. and von Hippel, P. H. (1987) *J. Mol. Biol.* 193, 723–750
25 Berg, O. G. and von Hippel, P. H. (1988) *J. Mol. Biol.* 200, 709–723
26 Heumann, J. M., Lapedes, A. S. and Stormo, G. D. (1994) in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 188–194, AAAI Press
27 Schneider, T. D., Stormo, G. D., Gold, L. and Ehrenfeucht, A. (1986) *J. Mol. Biol.* 188, 415–431
28 Mulligan, M. E., Hawley, D. K., Entriken, R. and McClure, W. R. (1984) *Nucleic Acids Res.* 12, 789–800
29 McClure, W. R. (1985) *Annu. Rev. Biochem.* 54, 171–204
30 Ringquist, S. *et al.* (1992) *Mol. Microbiol.* 6, 1219–1229