Is there a code for protein–DNA recognition? Probab(ilistical)ly...

Panayiotis V. Benos,[1] Alan S. Lapedes,[2] and Gary D. Stormo[1]*

Summary

Transcriptional regulation of all genes is initiated by the specific binding of regulatory proteins called *transcription factors* to specific sites on DNA called *promoter regions*. Transcription factors employ a variety of mechanisms to recognise their DNA target sites. In the last few decades, attempts have been made to describe these mechanisms by general sets of rules and associated models. We give an overview of these models, starting with a historical review of the somewhat controversial issue of a "recognition code" governing protein–DNA interaction. We then present a probabilistic framework in which advantages and disadvantages of various models can be discussed. Finally, we conclude that simplifying assumptions about additivity of interactions are sufficiently justified in many situations (and can be suitably extended in other situations) to allow a unifying concept of a "probabilistic code" for protein–DNA recognition to be defined. *BioEssays* 24:466–475, 2002. © 2002 Wiley Periodicals, Inc.

Introduction: some notes of historical interest

How do transcription factors recognise their DNA target sites? 25 years have elapsed since the work of Seeman et al.,[1] which initiated the study of the molecular basis of this recognition. Upon analysing the stereochemical properties of the residues, they noticed that two or more hydrogen bonds are required for the efficient discrimination between DNA bases by certain amino acids. In particular, they predicted that *Arg*, if placed appropriately, could specifically recognise guanine; and, similarly, *Asn* or *Gln* could recognise adenine. Interestingly, *Arg* ⇔ *G* is the most common contact found in the crystal structures today.[2]

In later years, the crystal structures of *Cro*[3] and *cI*[4] repressors of bacteriophage lambda and the CAP protein of *E. coli*[5] provided clues as to what the molecular mechanisms of the interactions could be and raised hopes that a simple set of rules might exist in nature that can adequately explain those interactions. In 1984, Pabo and Sauer[6] first proposed the term "Recognition Code" to describe such rules. However, they noted that, unlike the genetic code, the protein–DNA "recognition code" is degenerate in both directions. In other words, each DNA base could be recognised by a limited variety of amino acids and vice versa. In their review article, they modelled an additional possible contact: *Ser* ⇔ *A* with two hydrogen bonds (although, serine can act as either hydrogen-bond donor or acceptor; thus, its contribution to base specificity is more limited, Ref. 7).

Once the crystal structures of a handful of protein–DNA complexes became available, it was realised that transcription factors employ a variety of strategies to recognise their target sequences. That led Matthews to conclude that there is "no code for [protein–DNA] recognition" ;[8] or more specifically, that no *deterministic* protein–DNA recognition code, in analogy to the genetic code, exists. However, the determinism of the genetic code is exhibited in one way only, i.e., knowing the nucleotide triplets we can deduce the amino acid sequence. The reverse (from the amino acids to the nucleotide sequence) is probabilistic. A recognition code that is probabilistic in both directions, *P-code*, is supported by the work of several groups that have noticed clear base–amino acid preferences.[2,9–13]

Gene regulation is a complicated process. There are many important issues that deserve further investigation: protein–protein interactions and co-regulation of a gene by more than one transcription factor are merely two examples. However, this article focuses on the methods that have been used to model protein–DNA interactions and give some insights to some complicating factors, such as *energetic additivity*.

We start with a description of the problem of DNA specificity from the thermodynamic point of view. Then we describe the weight matrix methodology that has been widely used in modelling of binding sites and protein–DNA interactions. The problem of identification of transcription factor binding sites serves as an example to introduce the reader to basic aspects of weight matrix modelling. After describing the modelling of DNA-binding sites (of a "fixed" protein) via weight matrices, we extend this notion to include variation at the protein level. This leads us to a two-way table that associates bases and amino acids. Essentially all methods of modelling protein–DNA interactions use similar tables. Their differences are mainly in the number of parameters that needed to be

[1] Department of Genetics, Washington University.
[2] Theoertical Division, Los Alamos National Laboratories and Santa Fe Institute, 1399 Hyde Park Road, Santa Fe.

Funding agency: NIH (Grant HG00249) to GDS. The research of ASL was supported by the Department of Energy under contract W-7405-ENG-36.

*Correspondence to: Gary D. Stormo, Department of Genetics, Campus Box 8232, Washington University, School of Medicine, St Louis, MO 63110, USA. E-mail: stormo@genetics.wustl.edu

DOI 10.1002/bies.10073
Published online in Wiley InterScience (www.interscience.wiley.com).

estimated (*data compression*) and the way that these are estimated. We describe the basic characteristics of the *qualitative* and three *quantitative* methods that are currently used for such modelling.

The thermodynamics of protein–DNA interactions

DNA recognition by a particular regulatory protein can be a complex, multistep process (for a review see Ref. 14). Nevertheless, without loss of generality, it can be viewed as a chemical process, in which the rate of the reaction is limited by the rate that the two components (i.e., the protein and the DNA site) are brought together via diffusion. Once brought together into a protein–DNA complex, the dissociation rate depends on the *affinity*, or chemical complementarity, between them. In general, the higher the number of favourable chemical contacts between them the lower the dissociation rate and higher the affinity.

If we denote H_i the *binding energy* for any given DNA site S_i, then the probability that the protein would be bound to S_i (at equilibrium) is given by the Boltzmann distribution[15]:

$$P_i = \frac{e^{-H_i}}{Z} \qquad (1)$$

where Z is the *partition function* and is defined as the sum of the *exp(−H_x)* over all possible sites S_x:

$$Z \equiv \sum_x e^{-H_x}$$

Note that, in equation 1, we omitted the temperature factor $k_B T$ from the denominator of the energetic exponential. By doing this we assume that the temperature is constant throughout the binding process and that what we call "energy values" are expressed in ($k_B T$) units.

The *average binding energy* for this protein over all sites S_i would be:

$$\langle H \rangle = \sum P_i \cdot H_i = -\sum_i (P_i \cdot \ln\ P_i) - \ln Z \qquad (2)$$

The sum on the right side of the equation 2 is called the *entropy* of the probability distribution.

Suppose that we measure the binding energy of a given protein to all its possible DNA targets. These energies can form a probability distribution, P, via equation 1. Now let us assume that, instead of measuring directly all these energies, we estimated their values from a small set of measurements according to certain rules. Then the estimated values would form another probability distribution, Q, which generally would differ from P. A very useful measure of difference between two probability distributions is the *relative entropy*, which is defined as:

$$\mathcal{H}(P, Q) \equiv -\sum_i P_i \cdot \ln \frac{P_i}{Q_i}$$

If the two distributions, P and Q, are similar in the high probability states then their relative entropy is close to zero.

This formalisation makes clear the relation between the interaction energies and the specificity of a protein to certain DNA targets ("DNA recognition"). The lower the interaction energies are, the higher the probability that the particular target will be selected. It is also clear from this formalism that only the difference in energies matters to the probability of binding. Adding a constant term to all energies does not change the binding probability of any sequence (equation 1). In fact, we are free to choose the base line of energy such that *ln* $Z = 0$, in which case the average energy is equal to the entropy of the system, a classic result from thermodynamics. Moreover, the probability of equation 1 can also be viewed as a reflection of the preferences for certain targets by this protein; the higher the probability, the more frequently this target will be selected compared to the other targets. This idea provided the basis for the use of *weight matrices* in the study of various biological phenomena, including the modelling of regulatory regions and protein DNA interactions. Therefore, we present weight matrices in more detail in the next section.

Equation 1 is a simplified version of the one presented in Benos et al.[16] in that it does not include the relative frequencies of the different binding site types. However, these frequencies needed not be included if the different target site types are equiprobable or if S_i enumerates all possible DNA sites (instead of *types* of DNA sites). From a genomic point of view, the distinction is clear: two genes might have identical DNA binding sites, but we are interested to know when a particular one of them is bound.

It is important to realise the difference between *affinity* and *specificity*. The former is determined by chemical complementarity that a protein shows to a piece of DNA. This complementarity exists even in non-specific binding. In contrast, specificity is a measure of selective binding of the protein to "preferred" DNA targets. It is determined by the difference in affinity to different sequences. Hydrogen bonding is the major element in this sequence-specific recognition, since both bases and amino acids have hydrogen donor/acceptor potentials. Water-mediated bonds and van der Waals interactions also play a very important role in DNA binding. They are estimated to constitute approximately 15% and 65% of all contacts, respectively. However, their importance lies more on the stabilisation of the protein–DNA complex rather than in the specific recognition of the DNA target.[13] Finally, hydrophobic interactions and DNA conformation also contribute to the binding. But, although they can be sequence dependent (to some extent), they are also not sufficiently limiting by themselves for specific DNA recognition.[17]

Weight matrices in modelling

Weight matrices are very popular in the modelling of various biological phenomena. In the study of regulatory regions, for example, they were first employed as a method to distinguish functional binding sites from "non-functional" (but similar) ones in genomic DNA.[18] Later, they were put in a more statistical framework, having their elements derived from binding site examples.[19–21] The relationships between the *probability of occurrence* of each base in a particular position, its *energetic contribution* to the total binding energy and the *information content* of the binding sites were explored further in the following years (reviewed in Ref. 22). This general model had been the basis for various algorithms that try to identify regulatory sites in the promoter regions of coregulated genes.[23–26]

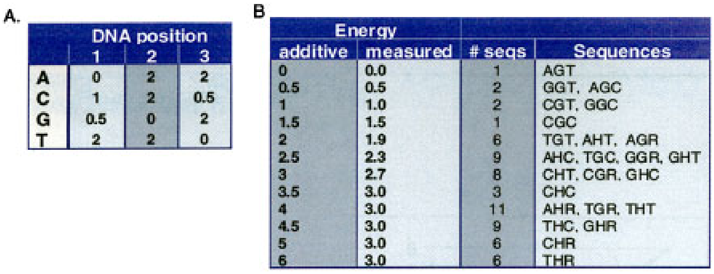
Although the identification of regulatory regions in genomic regions is not the main topic of this article, we use it as an example to introduce the reader to some important general concepts of modelling via weight matrices. Figure 1A shows a weight matrix for the DNA-binding site of a hypothetical protein. This toy binding site is 3 bp long and each column contains the *relative energetic potentials* for each of the four bases in the corresponding position. These values can be experimentally measured, but this usually requires laborious experiments in which the target DNA positions are systematically varied to all four bases (one position at a time; see below). Alternatively, one can estimate these parameters from a set of known (aligned) binding sites. In this case, the *binding energy* of every residue in each base position is estimated by the logarithm of its frequency, normalised with respect to the maximum frequency for this position. Hence, the base with the highest frequency has a binding energy of zero and all the rest have a positive value (see also Section "The thermodynamics of protein–DNA interactions"). Note that the convention that we adopt assigns lower energy values to stronger binding. Thus, the "consensus" binding site in our hypothetical example is AGT (Fig. 1).

Selecting nucleotides: does additivity hold?

A common feature of the regulatory region models (as well as the ones of the protein–DNA interactions) is the assumption that all base–amino acid contacts are contributing independently and, therefore, *additively* to the total *binding energy*. Equivalently, the *total probability* of the binding site is simply the product of the binding probabilities in each position (see also equation 1). Assuming that additivity holds, one can calculate the total binding energy for each of the 3 bp long binding sites. The corresponding values are presented in the first column of Fig. 1B. The complete list of measured energies requires a $4^3 = 64$ long table that was "compressed" into the 4×3 elements of Fig. 1A. This compression is only accurate if the additivity assumption is valid. We know that additivity is not exactly true. Thus, the additive model comprises an approximation of the true binding energies of a protein to different DNA sequences. How good an approximation it is depends on the protein that is being modelled. And how useful it is depends on what use is made of it. For many proteins, this model appears to be accurate enough for the prediction of binding sites in genomic DNA sequences.[27,28]

To demonstrate what effect non-additive interactions could have in the study of protein–DNA recognition, let us consider that the energy data presented in Fig. 1A were determined experimentally for each base in each position. This can be done, for example, by systematically varying each position to all four bases, whereas the other two positions are held fixed to the preferred ones (i.e., *A, G* and *T* for positions 1, 2 and 3, respectively). Similar experiments have been performed in the past.[29–31] If additivity holds, then from the data of Fig. 1A one could calculate the binding energy of every possible 3 bp long target sequence, by simply summing the energies associated with each base at each position. The result is shown in the first column of Fig. 1B ("additive energy").

In reality, however, the true binding energy tends to have an "upper bound", which is usually derived from the non-specific contacts (mostly to the DNA backbone). In other words, even in complete lack of chemical specific complementarity to the base pairs, the protein still displays some affinity to the DNA target. Let us assume that, in our example, this upper bound of the binding energy is 3.0 and the real binding energies (i.e., measured directly for each of the 64 DNA targets) are the ones presented in column 2 of Fig. 1B ("measured energy").



A.

	DNA position		
	1	2	3
A	0	2	2
C	1	2	0.5
G	0.5	0	2
T	2	2	0

B

Energy			
additive	measured	# seqs	Sequences
0	0.0	1	AGT
0.5	0.5	2	GGT, AGC
1	1.0	2	CGT, GGC
1.5	1.5	1	CGC
2	1.9	6	TGT, AHT, AGR
2.5	2.3	9	AHC, TGC, GGR, GHT
3	2.7	8	CHT, CGR, GHC
3.5	3.0	3	CHC
4	3.0	11	AHR, TGR, THT
4.5	3.0	9	THC, GHR
5	3.0	6	CHR
6	3.0	6	THR

Figure 1. A: The binding energy matrix for a hypothetical protein. In this example, the DNA target size is 3 bases long. The numbers are usually calculated from a set of aligned binding sites and correspond to the logarithm of the normalised frequencies. **B:** Example of relative energy values of the interactions between the hypothetical protein and all its 3 bp long DNA targets. The first column presents the energy values that the protein has towards all possible 3-bp long targets under *additivity*. These values have been calculated from the data of **A**. The second column contains "experimentally" defined energy values. This hypothetical example shows a case where additivity holds only on the lower energy (i.e., higher affinity) states. In this table, *R* denotes *A* or *G* and *H* denotes *A* or *C* or *T*.

Figure 2 illustrates one of the ways that the additivity is violated. The predicted energy values have been plotted against themselves and against the observed ones. The straight line represents the additive model and the curved is derived from the measured energies of our hypothetical example. This example shows a linear region and a plateau, not unlike many real proteins.(29,32) Additivity holds for the low energy sites, those with energies most similar to the preferred sequence, but has a non-specific energy that tends to 3.0, for triplets with predicted energies greater than 1.5. This means that over 90% of the predicted energies are wrong and many of them significantly so, by up to 3.0 units.

Is the additive model (presented as the straight-line graph in Fig. 2) a good approximation for this protein? By conventional "goodness-of-fit" criteria, it may not look like a good approximation, because many of the points deviate far from the straight line. Based on this observation, one may conclude that there is no additive model, of whatever slope, that will be a good fit to the experimental data. However, in our opinion, this is not the most appropriate criterion to use. There are three other, perhaps more suitable, criteria that could be used (they are also closely related to each other): (1) the difference in observed and predicted *free energy* of the system, (2) the probability that the preferred sites are bound, and (3) the concentration of protein necessary to achieve a desired level of saturation of the preferred sites.

Note that, for our hypothetical example, the *entropy* for the probability distribution of the "real" (i.e., measured) values is 3.65 (as defined by equation 2 with *ln* $Z=0$), whereas the additivity model gives 3.19. The *relative entropy* of the two distributions is 0.18. But we can find an even better additive model that minimises the difference between the two probability distributions (*best additive model*). This model has an *entropy* of 3.71, with *relative entropy* only 0.05.

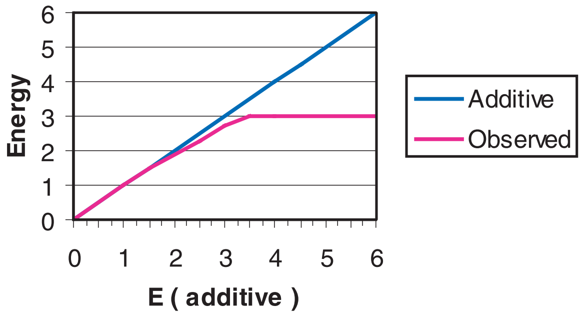


Figure 2. Graph of the experimentally measured energy values (curved line) and the ones calculated from Fig. 1A under additivity (straight line). The plot is based on a hypothetical example.

This is really a quite good fit to the real data, which can be better appreciated by the other criteria mentioned above. The true probability of the protein binding to the preferred site is 13.8%, and the best additive model predicts 10.4%. If we assume that the DNA targets are equally randomised, then the number of protein molecules needed to have the preferred site 95% saturated can be estimated by simulation experiments. This number is 16 for the "measured" values of Fig. 1B and Ref. 20 for the best additive model. Similar calculations based on the Mnt protein are provided in Ref. 31 (The ideas described here also hold in the case of biased representation of the DNA targets, but with slightly more complicated formulas(16)).

The point of this exercise is to indicate that protein–DNA interactions may be far from additive overall, but still be well represented by an additive model, provided that the non-additivity shows up mainly in the low-affinity sites. If non-additivity shows up strongly among the high-affinity sites then the additive model will probably be inadequate for most purposes. In such cases, more-complex models can be used. For instance, instead of assigning an energy value for each base at each position, we could use a model that assigns an energy value for each di-nucleotide. If the non-additivity were due primarily to interactions between adjacent bases, then this model would accurately capture those contributions and provide a representative, additive model.(33,34)

The important issue in deciding whether an additive model is appropriate is the trade off between the loss of information, as measured by the difference in true and predicted average binding energies, and the reduction in the number of parameters (i.e., *compression*) of the representation. In the hypothetical protein example, the true binding energies comprise 64 parameters (or, rather, 63, since they are relative to the preferred site). The model using the additive energy matrix has only 12 parameters (really just 9). This is a large compression in the representation, which costs a relatively small amount of accuracy in the prediction of the behaviour of the system. A protein that binds to *l*-long sequences has 4^l binding energies, but an additive model has only $4 \times l$ parameters. As long as the error is small, this is a very "economical" representation.

Selecting proteins

In the above example, the protein was considered fixed and measurements were made to all possible DNA-binding sites. The converse is also doable, although more laborious; therefore, fewer examples exist in the literature.(35–37) In these cases, binding energy measurements are performed for every possible protein sequence to a particular (fixed) DNA target. However, usually the only part of the protein that is randomised are the amino acid positions involved in sequence-specific binding. The rest of the protein is usually held constant, so that it folds properly and interacts with the DNA backbone in the same way as the wild-type one. The amino acid positions that vary are the ones that interact directly with the DNA bases, typically a small number (about the same number as the bases in the target site).

If we consider that, in the example of the previous paragraph, the protein uses three amino acids for DNA recognition, then the list of all possible proteins (with respect to these amino acid positions) would be "just" $20^3 = 8{,}000$. Although this is a large number, it is enormously smaller than all possible protein sequences of typical lengths. If we employ the same additive model that we did for the variable DNA to these experiments, we will have to estimate only $20 \times l$ parameters, where l is the number of "contacting" amino acid positions. In our example, this number would be *60*.

However, what is not known is how good a representation is the additive model for the amino acids. Because of the interactions between the amino acid side chains, the amino acid positions might be less independent than the DNA ones. At this time, there is not enough data to answer this question in general. And as before, the criteria to be used are the trade-off between the accuracy and the compression in the representation. If non-additivity shows up mainly for the low-affinity sequences, we may still get good additive models. If this is not the case, then more complex models may be needed, but still employing significantly less parameters than the number of all possible sequences. For example, the number of parameters for a di-amino acids model over a 3-residue "protein" is $20^2 \times 2 = 800$, while the total number of possible sequences is *8,000*. Longer proteins result in greater compression savings. However, we do not know *a priori* what order of compression is needed.

Modelling interactions both ways

In the previous paragraphs, we described the general features of additive, energetic/probabilistic models of DNA–protein interactions where one component, either the DNA or the protein, is held fixed and the other varies. We can easily extend the model to "two dimensions" where both the DNA and protein are variable. In fact, the full interaction of all proteins of a particular type to all possible DNA targets can be described in terms of a gigantic matrix that contains the binding energy of every protein sequence to every DNA sequence. For our hypothetical example (three amino acids contact three bases), this would mean a $4^3 \times 20^3$ table of binding energies, which represents a number of experiments that is easier said than done. Employing the additivity assumption causes a simplification to the problem at the expense of some loss of information. Indeed, in such a case, one needs to estimate "only" $(4 \times 3) \times (20 \times 3)$ parameters. That is, of course, if one considers all possible interactions between the three amino acids and the three bases. If the mode of interactions is the "one-to-one" (i.e., each amino acid interacts with one base position only and vice versa), then the parameters in the model are only $3 \times 4 \times 20$. It remains to be determined how accurate such a model can be.

From this point of view, it is clear that the additivity assumption represents a compression of the information in the data. In other words, we lose some information in order to reduce the number of parameters that we need to estimate. In the following sections, we will describe methods that have been developed to estimate different numbers of parameters in similar weight matrices.

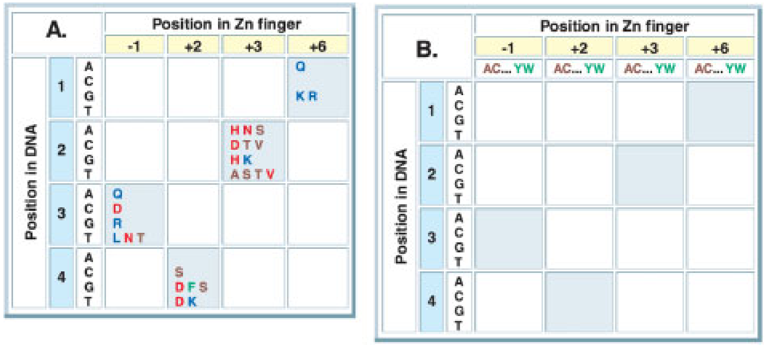
Finally, we must note another assumption/simplification that one silently makes, when protein–DNA interactions are modelled in this way: i.e., the amino acid substitutions in a protein do not alter the bonding "recognition scheme" to the DNA. Or, they do so only to the extent that the specificity of the protein is maintained. But this is not exactly true always. For example, it has been reported that, in some cases, modest changes in the docking arrangement are utilised to accommodate new side chain–base and side chain–phosphate interactions.[7] It is not known how much these "docking rearrangements" alter the overall binding specificity. We know, for example, that the vast majority of the protein–DNA contacts are non-specific ones, mainly to the DNA backbone.[13] So, it might be that this is a similar situation to the deviation from additivity that we described before. Nevertheless, this phenomenon is expected to be family specific. Thus, we might end up modelling a protein family with multiple "codes", depending on the residues present in some key amino acid positions. This is a very interesting subject that is open for further investigation.

Estimating the parameters

During the last 10 years, there have been a number of attempts to model protein–DNA interactions. All of the models developed so far have focused on additive interactions. There are many reasons for making this choice. A key reason is that currently available data are far from enough to model putative non-additive interactions accurately. Also, in the previous section we explained why, for modelling purposes, it is only important that the additivity assumption is approximately true for the high affinity binding sites.

Depending on the representation of the interactions, models can be classified into two types: qualitative and quantitative. We call a model *qualitative* when it represents the interactions in a *binary* way (i.e., a protein does or does not bind to a DNA target sequence). By contrast, a model is called *quantitative* if it provides a *measure* for the binding affinity.

Figure 3 presents an example of the two model types, in the form of weight matrices. Both models refer to the Early Growth Response (EGR) family of transcription factors (also known as Zif268, Krox, NGFI-A). The EGR proteins belong to the Cys2His2 zinc-finger protein family. They were originally identified in mammals, but homologous proteins have been cloned in a variety of species (including zebrafish and *Xenopus laevis*). Their DNA target is recognised via three



A.			Position in Zn finger			
			-1	+2	+3	+6
Position in DNA	1	A				Q
		C				
		G				K R
		T				
	2	A			H N S	
		C			D T V	
		G			H K	
		T			A S T V	
	3	A	Q			
		C	D			
		G	R			
		T	L N T			
	4	A				
		C		S		
		G		D F S		
		T		D K		

B.			Position in Zn finger			
			-1	+2	+3	+6
			AC... YW	AC... YW	AC... YW	AC... YW
Position in DNA	1	A C G T				▨
	2	A C G T			▨	
	3	A C G T	▨			
	4	A C G T		▨		

Figure 3. Schematic representation of the qualitative and the quantitative models for the EGR protein family. Each zinc finger employs four amino acids (positions −1, +2, +3 and +6 of the α-helix respectively) to recognise a 4 bp long target site. The qualitative model (**A**) consists of a look-up table that lists all observed base-amino acid contacts. On the other hand, a quantitative model (**B**) provides a score for each of those contacts (for simplicity, we have omitted the actual values from this table; see Fig. 4 for an example). Obviously, the qualitative model is a degenerate quantitative, with the scores being only *1*s or *0*s. The shaded submatrices correspond to the contacts that have been observed in the co-crystal structure of the EGR protein. The amino acids in this figure are colour coded according to size (see also next figure). The qualitative model has been adapted from Refs 40,43. In the recent years, it has been used to predict and explain interaction data of this family.[43] The amino acids referred in this figure are colour coded according to size/properties: *brown*, *red*, *blue* and *green* represent small, medium, large and aromatic amino acids, respectively.

highly conserved α-helices. The structure of the three zinc-fingers of the protein bound to the consensus DNA sequence was initially solved crystallographically at 2.1Å[38] and subsequently refined to 1.6Å.[39] The target site is 10 bp long, where each finger contacts four of these bases (with one base overlap between the fingers).[39] The topology of the molecules in the solved crystal structure showed that the four "critical" amino acids of a finger could contact one base each on the target site ("one-to-one" mode of interactions). The shaded areas in Fig. 3 represent the contacting scheme. If we lacked the crystal structure, a complete model should include the additional possible contacts (white submatrices in Fig. 3; "all-to-all" mode of interactions).

The basic difference between qualitative and quantitative models is that the former merely consist of a list of all observed contacts, whereas the latter provide a measure of the affinity of these contacts. As is obvious from Fig. 3, any quantitative model can be transformed into qualitative by setting a threshold to separate "binding" from "no binding". Various quantitative models differ in the way that they determine the affinity of the contacts (i.e., parameter estimation).

Qualitative modelling

In 1992, Desjarlais and Berg[10] studied the DNA specificity of a number of *Sp1*-derived proteins and they organised their data into a "set of rules". Using their table, one could predict the amino acids of a putative *Sp1*-derived protein that would bind with high specificity a given tri-nucleotide of the form *GNK* (N:*A* or *C* or *G* or *T*;K:*G* or *T*) and vice versa. Although they performed some quantitative experiments, they did not incorporate them into their model, thus their model remained qualitative.

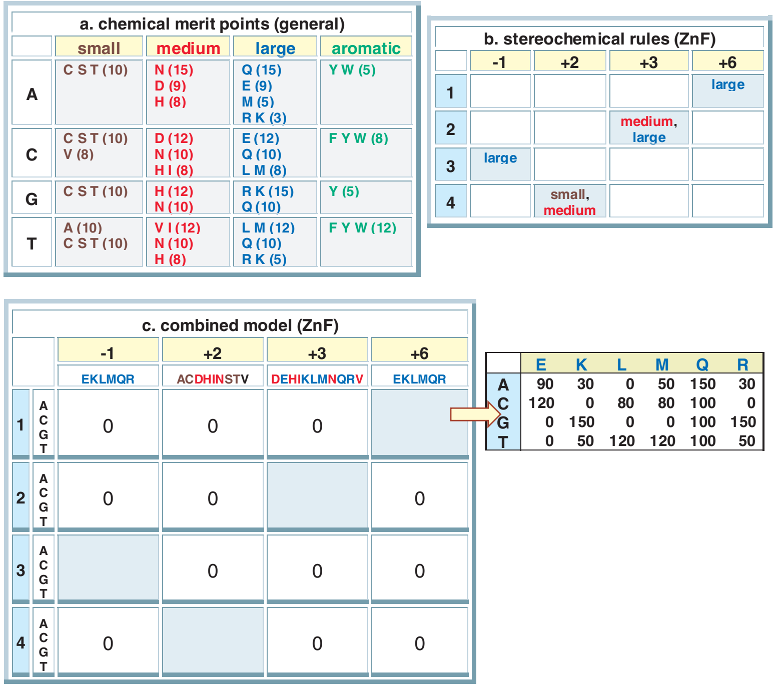
More recently, Choo and Klug initially[40,41] and Pabo and colleagues later[42,43] proposed a qualitative model to explain/predict the DNA recognition for members of the EGR protein family. An example of the *qualitative model* proposed by the two groups is presented in Fig. 3A. It consists of a list of all amino acids that have been found to contact particular bases and it is position specific. For example, lysine at positions +3 and +6 has been found to contact guanine (at the corresponding nucleotide positions 2 and 1); whereas lysine at position +2 prefers thymine (at nucleotide position 4).

Since the tabulated base–amino acid interaction data are position-specific, this model has inherent both chemical and stereochemical properties of the protein and the DNA (see also Section "Chemical and stereochemical rules"). In fact, Wolfe et al.[43] used it to predict DNA-binding specificities of three *EGR*-derived proteins. In each of the three cases, the model predicted correctly six out of nine contacts.

The disadvantage of a qualitative model is, of course, that it cannot make quantitative predictions. For instance, in the example of Fig. 3, does *Asp* in positions +2, −1 and +3 bind with the same strength the corresponding bases? A limitation of this method is its dependence on structural as well as other experimental data for making efficient predictions.

Chemical and stereochemical rules

Suzuki and his co-workers have proposed a method to model the protein–DNA interactions in a quantitative way, based on a set of chemical and stereochemical rules that they presented.[27,28] According to that model, a *chemical merit point* is assigned for each permissible base–amino acid contact. These points are defined in a semi-arbitrary way from the chemical properties of the residues. Similarly, a *stereochemical merit point* is assigned for each base–amino acid contact position, based on information from the co-crystal structures and the size of amino acids (Fig. 4). The *stereochemical merit points* have the values of 5 or 10, depending on the type of



a. chemical merit points (general)				
	small	medium	large	aromatic
A	C S T (10)	N (15) D (9) H (8)	Q (15) E (9) M (5) R K (3)	Y W (5)
C	C S T (10) V (8)	D (12) N (10) H I (8)	E (12) Q (10) L M (8)	F Y W (8)
G	C S T (10)	H (12) N (10)	R K (15) Q (10)	Y (5)
T	A (10) C S T (10)	V I (12) N (10) H (8)	L M (12) Q (10) R K (5)	F Y W (12)

b. stereochemical rules (ZnF)				
	-1	+2	+3	+6
1				large
2			medium, large	
3	large			
4		small, medium		

c. combined model (ZnF)					
		-1	+2	+3	+6
		EKLMQR	ACDHINSTV	DEHIKLMNQRV	EKLMQR
1	A C G T	0	0	0	
2	A C G T	0	0		0
3	A C G T		0	0	0
4	A C G T	0		0	0

	E	K	L	M	Q	R
A	90	30	0	50	150	30
C	120	0	80	80	100	0
G	0	150	0	0	100	150
T	0	50	120	120	100	50

Figure 4. Schematic representation of the Suzuki's quantitative model for the ZnF family. The scheme is adapted from Ref. 27. According to the model, *the chemical merit points* (a) and the stereochemical rules for the EGR family (b) are combined to give the "recognition code" table (c). The chemical merit points are based solely on the physicochemical properties of the molecules, hence they are family/position independent. The stereochemical merit points are family-specific and depend on the crystallographic data. For the EGR protein family they assign the value of 10 for each of the contact/amino acid group presented in (b) and zero otherwise. For example, for the position +6 of the helix, the large amino acids (*E*, *K*, *L*, *M*, *Q* and *R*) are granted 10 *stereochemical merit points* for contacting base at position 1 of the DNA target. All other amino acids have zero stereochemical merit points for this contact. The contacts of amino acid position +6 to all other base positions have zero stereochemical points too. The combination of the two matrices (a) and (b) is represented in matrix (c). As an example, we provide the combined score of the base-amino acid contacts for the large amino acids in the adjacent table. The model proposed by Suzuki et al.[27] assigns a score to each protein–DNA pair. This score is the sum of values from (c) for all "valid" contacts. In the original paper the authors provided the stereochemical rules for other protein families as well (*HTH*, *PH* and *C4*). The amino acids referred in this figure are colour coded according to size/properties: *brown*, *red*, *blue* and *green* represent small, medium, large and aromatic amino acids, respectively.

contact. For the EGR protein family, the stereochemical merit points have a value of 10 for all "valid" contacts (i.e., the ones derived from the crystal structure) on all "permissible" amino acid groups (see Fig. 4b) and zero otherwise.

For a given protein from the families that are modelled in this way, a score is assigned to every DNA target. This score is the sum of the products: *(chemical merit point)* × *(stereochemical merit point)*, over all base–amino acid contacts (essentially, the sum of the weights from Fig. 4c). The higher the score, the higher should be the affinity of this protein to the DNA.

This model was tested on data from various families of transcription factors (i.e., HTH, PH, ZnF, C4, etc). For the evaluation of their method, the authors introduced the *specificity index*, which is defined as:

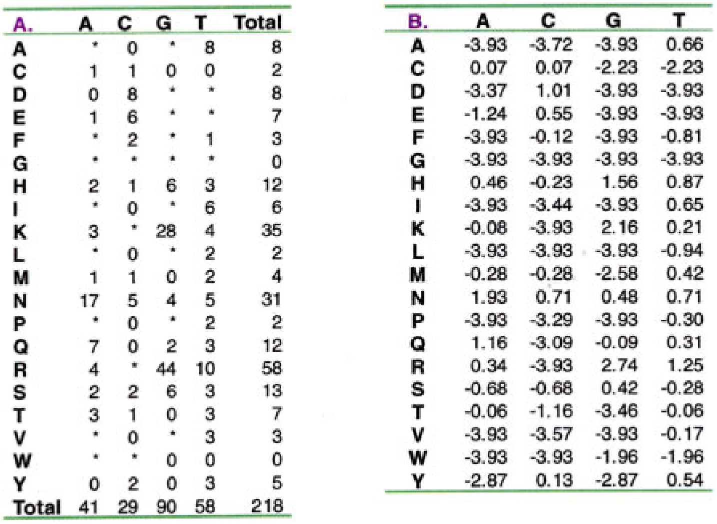
$$SI \equiv 100 - n - (m/2)$$

where n and m are the percentages of the DNA sequences that score higher than and equal to the real binding sequence, respectively. If the topmost prediction is the real (and unique) binding sequence, then the specificity index would be 100. The average specificity indices, calculated for the known binding sites of the tested protein families, were between 92 (for HTH) and 99 (for C4).

As innovative as it is, this method has two limitations: (a) both the "chemical" and the "stereochemical merit points" (i.e., the parameters of the model) have been assigned semi-arbitrarily (although, to some extent, they could be determined experimentally) and (b) in order to determine the "stereochemical merit points", one needs to know the co-crystal structure of at least one member of the protein family and its DNA target.

Frequency-based modelling:exploiting structural data

Following a different approach, the group of Margalit developed another quantitative model for protein–DNA interactions. This model is based on the frequencies of the contacts found in co-crystal structures.[2,44] For each possible base–amino acid contact, they assigned a score $S_{ij} = ln[f_{ij}/(f_i \times f_j)]$, where f_{ij} is the observed frequency of contacts between the amino acid i and the base j, f_i was set to be the frequency of amino acid i in the SWISS-PROT protein database and f_j was 0.25 for all bases. These values form a 20 × 4 table, which was



A.	A	C	G	T	Total
A	*	0	*	8	8
C	1	1	0	0	2
D	0	8	*	*	8
E	1	6	*	*	7
F	*	2	*	1	3
G	*	*	*	*	0
H	2	1	6	3	12
I	*	0	*	6	6
K	3	*	28	4	35
L	*	0	*	2	2
M	1	1	0	2	4
N	17	5	4	5	31
P	*	0	*	2	2
Q	7	0	2	3	12
R	4	*	44	10	58
S	2	2	6	3	13
T	3	1	0	3	7
V	*	0	*	3	3
W	*	*	0	0	0
Y	0	2	0	3	5
Total	41	29	90	58	218

B.	A	C	G	T
A	-3.93	-3.72	-3.93	0.66
C	0.07	0.07	-2.23	-2.23
D	-3.37	1.01	-3.93	-3.93
E	-1.24	0.55	-3.93	-3.93
F	-3.93	-0.12	-3.93	-0.81
G	-3.93	-3.93	-3.93	-3.93
H	0.46	-0.23	1.56	0.87
I	-3.93	-3.44	-3.93	0.65
K	-0.08	-3.93	2.16	0.21
L	-3.93	-3.93	-3.93	-0.94
M	-0.28	-0.28	-2.58	0.42
N	1.93	0.71	0.48	0.71
P	-3.93	-3.29	-3.93	-0.30
Q	1.16	-3.09	-0.09	0.31
R	0.34	-3.93	2.74	1.25
S	-0.68	-0.68	0.42	-0.28
T	-0.06	-1.16	-3.46	-0.06
V	-3.93	-3.57	-3.93	-0.17
W	-3.93	-3.93	-1.96	-1.96
Y	-2.87	0.13	-2.87	0.54

Figure 5. A weight matrix for position-independent modelling of protein–DNA interactions. This model was developed by Margalit's group.[2,44] Based on the frequencies of the particular base-amino acid contacts that were observed in the crystal structures (**A**) scores are assigned (**B**). Subsequently, these scores are used for ranking different protein–DNA pairs. Contacts marked with the star symbol (*) are the ones that exhibit no chemical complementarity; thus, according to the authors, they cannot be found in nature. A lowest score of −3.93 was assigned to these contacts. By contrast, contacts with zero frequency are those that can be observed in nature; but no examples were found in the particular set of crystal structures. This model provides scores for the contacts in a family independent way. In the case of EGR family for example, all four weight submatrices (Fig. 3) will have the same values (B).

used to calculate the total score for a particular protein–DNA interaction, assuming additivity (Fig. 5).

On predicting DNA-binding experimental data of the *EGR* protein family, this method performed satisfactorily.[2] Using the SELEX data (see below) provided in two studies,[40,45] their algorithm was able to rank approx. 50% of the experimentally selected triplets in position 6 or higher (i.e. to the top 10% of all the 64 triplets). Moreover, the calculated scores for various protein–DNA pairs and the experimentally assessed relative free energies in the same two studies were shown to be correlated ($R = -0.\ 79$ and -0.49 for the two studies, respectively). The negative sign (i.e., *anti*-correlation) is expected, since the convention in Margalit's model is that a high (positive) score of a particular protein–DNA pair denotes high affinity of the two, which corresponds to lower energy values. Finally, in a more recent report from this group,[46] an expanded set of hydrogen bonds was used to calculate the scores thus enhancing the performance of their algorithm.

Nevertheless, there are some limitations inherent to this approach. For example, it assumes that the base–amino acid contacts are position independent. That is, all amino acid positions are treated as equivalent, with respect to the base–amino acid contact preferences. Although, the chemical properties of the base–amino acid contacts are position-independent, the same is not true for the stereochemical properties.[28] This represents a further compression of the data, since the model assigns to them essentially an average over all contacts contained in the crystal structures of the training set. Thus, treating all base–amino acid contacts as equivalent causes a loss of information (due to averaging over all contacts).

Another drawback of the method is that it can only model the "one-to-one" type of interactions (i.e., one amino acid contacts one base and vice versa); which is not always the case.[47] In fact, in a recent study of protein–DNA interactions at the atomic level,[13] Thornton's group reported as many as 43 examples of complex interactions found in the crystal structures. This constitutes 12% of all contacts between amino acids and bases in their data set. Finally, the small size of the "training set" used in this model (53 crystal structures with 218 contacts) imposes a limitation on its prediction capabilities.

Frequency-based modelling:exploiting selection data

Recently, our group presented another quantitative approach.[16] Based on the statistical mechanics theory that was briefly described earlier (Section "The thermodynamics of protein–DNA interactions"), we developed an algorithm for modelling DNA–protein interactions. This algorithm is named SAMIE (Statistical Algorithm for Modelling Interaction Energies) and uses data from selection experiments to estimate the parameters of the model, for any given protein family. There are, generally, two types of selection experiments it can utilise: SELEX and phage display.[40,48] The idea for *SELEX* experimentation is that a particular ("fixed") protein selects a number of DNA targets from a pool of randomised oligonucleotides. Subsequently, mutants of the protein are used in the same way. The reverse procedure ("fixed" DNA, randomised protein) is called *phage display*. The randomised counterparts (DNA or proteins, respectively) that are recovered from these experiments form high-affinity interactions with the fixed ones. However, it is known that sometimes the highest affinity ones might be missed, due to the stochastic nature of these processes. Nevertheless, the higher affinity sites should have a higher probability of being selected. SAMIE is essentially a maximum likelihood method. It exploits data from SELEX and/or phage display experiments (individually or in a combined set) to estimate the energetic potentials between the different residues in all base–amino acid contacts of the modelled protein. This estimation is achieved *via* a maximisation of the log-likelihood of the data (based on the statistical mechanics theory). There are no restrictions as to what the contacting scheme should be (e.g., "one-to-one", "one-to-many", "many-to-many", etc).

In our previous study,[16] the SAMIE algorithm was trained on SELEX data from the EGR protein family and it was then able to predict DNA-binding sites of *EGR*-derived proteins that were not included in the training set, as well as the SELEX results of proteins belonging to the MIG family (a yeast family of Cys_2His_2 zinc-finger transcription factors that bears no similarity to EGR outside the finger regions). Furthermore, the predicted energy values coincide well ($R = 0.8$) with experimental data from Ref. 49.

SAMIE is a fairly general model, as it requires no prior detailed knowledge of a specific contacting scheme (e.g., a co-crystal structure). However, should this knowledge be available, it reduces significantly the number of parameters that need to be estimated. For example, the model of the *EGR* protein family (Fig. 3) requires $4 \times 20 \times 16 = 1,280$ parameters if all interactions are considered ("all-to-all" model), but this number is reduced $4 \times 20 \times 4 = 320$ for the contacts that have been observed in the co-crystal structure ("one-to-one" model). SAMIE's only assumption is the *additivity* of the energetic contributions. Extensions to SAMIE can encompass non-additivity of individual base–amino interactions by, for example, consideration of di-nucleotides and di-amino acids.

The idea behind this algorithm is that if a protein binds to a DNA with high affinity, this should be reflected in the observed frequencies of their base–amino acid contacts. This is essentially the same idea behind the model presented by the Margalit's group.[2,46] However, SAMIE also takes into consideration the variation in preferences due to the topology of the binding. This consideration is close to the *stereochemical rules*, included in the model of Suzuki and colleagues.[27,28]

The exploitation of statistical mechanics theory for the calculation of the individual contact "energy" values gives SAMIE a strong theoretical basis. In fact, a model of this type must be a perfect representation of the interactions to some level, but limiting amounts of data constrain the number of parameters that can be estimated, forcing some approximations. In the case of modelling a single zinc finger of the Cys_2His_2 type, the *additivity* assumption reduces the number of parameters to 1,280 for the "all-to-all" and 320 for the "one-to-one" model, respectively. Thus, even under this assumption, a relatively high number of training examples is required for a complete, efficient training. Given that, it is notable that, with a training set of 675 training vectors derived from SELEX experiments the predictions are reasonably good.[16]

Conclusions

The search for a simple, deterministic recognition code had to be abandoned after the first few protein–DNA structures were solved. But it is also clear that amino acids and base pairs have preferred interacting partners which determine the probabilities of their combinations being used in regulatory proteins and their binding sites. Therefore a probabilistic recognition code may provide good predictions of high-affinity interactions. An important consideration is whether an additive model can be a good approximation. Clearly the contributions of individual amino acids and base pairs are not strictly additive, but we show that reasonable additivity need only hold for the high-affinity combinations in order to have a useful model. More complicated models are possible but require more parameters to be estimated (i.e., requiring more experimental data).

The EGR family of proteins has a more extensive collection of known protein–DNA interactions than any other family. Even so it is not enough to get good estimates of all the parameters in our model. But current methods such as SELEX and phage-display allow the rapid collection of more sequence data. Furthermore, quantitative affinity data can be utilized directly to enhance the parameter estimation. New methods for obtaining high throughput quantitative binding data will help enormously.[34,50]

One of the most interesting, open questions is how similar the probabilistic codes will be for different protein families. Can binding energies for other zinc finger families be accurately predicted using the EGR family model? Can the specificity of HTH proteins, which also use α-helices for recognition, be predicted from the same, or very similar models? These questions can only be addressed by further experiments. But if one can determine good probabilistic codes for all DNA-binding protein families, then one could predict the binding sites, and the set of regulated genes, for all regulatory proteins within any sequenced genome. That would constitute a major advance in our ability to understand and model entire regulatory networks.

Acknowledgments

The authors wish to thank the Santa Fe Institute, where part of this work was performed.

References

1. Seeman NC, Rosenberg JM, Rich A. Sequence-specific recognition of double helical nucleic acids by proteins. Proc Natl Acad Sci USA 1976;73:804–808.
2. Mandel-Gutfreund Y, Margalit H. Quantitative parameters for amino acid–base interaction:implications for prediction of protein–DNA binding sites. Nucleic Acids Res 1998;26:2306–2312.
3. Anderson WF, Ohlendorf DH, Takeda Y, Matthews BW. Structure of the cro repressor from bacteriophage lambda and its interaction with DNA. Nature 1981;290:754–758.
4. McKay DB, Steitz TA. Structure of catabolite gene activator protein at 2.9 Å resolution suggests binding to left-handed B-DNA. Nature 1981;290:744–749.
5. Pabo CO, Lewis M. The operator-binding domain of lambda repressor: structure and DNA recognition. Nature 1982;298:443–447.

6. Pabo CO, Sauer RT. Protein–DNA recognition. Annu Rev Biochem 1984;53:293–321.
7. Elrod-Erickson M, Benson TE, Pabo CO. High-resolution structures of variant Zif268–DNA complexes:implications for understanding zinc finger–DNA recognition. Structure 1998;6:451–464.
8. Matthews BW. Protein–DNA interaction. No code for recognition. Nature 1988;335:294–295.
9. Nardelli J, Gibson TJ, Vesque C, Charnay P. Base sequence discrimination by zinc-finger DNA-binding domains. Nature 1991;349:175–178.
10. Desjarlais JR, Berg JM. Toward rules relating zinc finger protein sequences and DNA-binding site preferences. Proc Natl Acad Sci USA 1992;89:7345–7349.
11. Jamieson AC. Antipersonnel mines: the global endemic. Ann R Coll Surg Engl 1996;78:396.
12. Wolfe SA, Nekludova L, Pabo CO. DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct 2000;29:183–212.
13. Luscombe NM, Laskowski RA, Thornton JM. Amino acid–base interactions:a three-dimensional analysis of protein–DNA interactions at an atomic level. Nucleic Acids Res 2001;29:2860–2874.
14. von Hippel PH, Berg OG. Facilitated target location in biological systems. J Biol Chem 1989;264:675–678.
15. Klotz IM. Introduction to Biomolecular Energetics. London UK: Academic Press Inc; 1986.
16. Benos PV, Lapedes AS, Fields DS, Stormo GD. SAMIE:statistical algorithm for modeling interaction energies. Pac Symp Biocomput 2001; 115–126.
17. Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins. Trends Biochem Sci 1988;13:207–211.
18. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res 1982;10:2997–3011.
19. Harr R, Haggstrom M, Gustafsson P. Search algorithm for pattern match analysis of nucleic acid sequences. Nucleic Acids Res 1983;11:2943–2957.
20. Staden R. Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res 1984;12:505–519.
21. Mulligan ME, Hawley DK, Entriken R, McClure WR. Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity. Nucleic Acids Res 1984;12:789–800.
22. Stormo GD. Consensus patterns in DNA. Methods Enzymol 1990;183: 211–221.
23. Stormo GD, Hartzell GW, 3rd. Identifying protein-binding sites from unaligned DNA fragments. Proc Natl Acad Sci USA 1989;86:1183–1187.
24. Lawrence CE, Reilly AA. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins 1990;7:41–51.
25. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals:a Gibbs sampling strategy for multiple alignment. Science 1993;262:208–214.
26. Heumann JM, Lapedes AS, Stormo GD. Neural networks for determining protein specificity and multiple alignment of binding sites. Proc Int Conf Intell Syst Mol Biol 1994;2:188–194.
27. Suzuki M, Yagi N. DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. Proc Natl Acad Sci USA 1994;91:12357–12361.
28. Suzuki M, Brenner SE, Gerstein M, Yagi N. DNA recognition code of transcription factors. Protein Eng 1995;8:319–328.
29. Takeda Y, Sarai A, Rivera VM. Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. Proc Natl Acad Sci USA 1989;86:439–443.
30. Sarai A, Takeda Y. Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. Proc Natl Acad Sci USA 1989;86:6513–6517.
31. Fields DS, He Y, Al-Uzri AY, Stormo GD. Quantitative specificity of the Mnt repressor. J Mol Biol 1997;271:178–194.
32. Frank DE, Saecker RM, Bond JP, Capp MW, Tsodikov OV, Melcher SE, Levandoski MM, Record MT Jr, Thermodynamics of the interactions of lac repressor with variants of the symmetric lac operator:effects of converting a consensus site to a non-specific site. J Mol Biol 1997;267: 1186–1206.
33. Stormo GD, Schneider TD, Gold L. Quantitative analysis of the relationship between nucleotide sequence and functional activity. Nucleic Acids Res 1986;14:6661–6679.
34. Man TK, Stormo GD. Non-independence of Mnt repressor–operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. Nucleic Acids Res 2001;29:2471–2478.
35. Lehming N, Sartorius J, Kisters-Woike B, von Wilcken-Bergmann B, Muller-Hill B. Mutant lac repressors with new specificities hint at rules for protein–DNA recognition. EMBO J 1990;9:615–621.
36. Rebar EJ, Pabo CO. Zinc finger phage:affinity selection of fingers with new DNA-binding specificities. Science 1994;263:671–673.
37. Miller JC, Pabo CO. Rearrangement of side-chains in *Zif268* mutant highlights the complexities of zinc finger–DNA recognition. J Mol Biol 2001;313:309–315.
38. Pavletich NP, Pabo CO. Zinc finger–DNA recognition: crystal structure of a Zif268-DNA complex at 2. 1 A. Science 1991;252:809–817.
39. Elrod-Erickson M, Rould MA, Nekludova L, Pabo CO. Zif268 protein–DNA complex refined at 1.6 Å: a model system for understanding zinc finger–DNA interactions. Structure 1996;4:1171–1180.
40. Choo Y, Klug A. Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. Proc Natl Acad Sci USA 1994;91:11168–11172.
41. Choo Y, Klug A. Physical basis of a protein–DNA recognition code. Curr Opin Struct Biol 1997;7:117–125.
42. Greisman HA, Pabo CO. A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. Science 1997;275:657–661.
43. Wolfe SA, Greisman HA, Ramm EI, Pabo CO. Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. J Mol Biol 1999;285:1917–1934.
44. Mandel-Gutfreund Y, Schueler O, Margalit H. Comprehensive analysis of hydrogen bonds in regulatory protein–DNA complexes: in search of common principles. J Mol Biol 1995;253:370–382.
45. Desjarlais JR, Berg JM. Length-encoded multiplex binding site determination: application to zinc finger proteins. Proc Natl Acad Sci USA 1994;91:11099–11103.
46. Mandel-Gutfreund Y, Baron A, Margalit H. A structure-based approach for prediction of protein binding sites in gene upstream regions. Pac Symp Biocomput 2001;139–150.
47. Desjarlais, Berg JM. Redesigning the DNA-binding specificity of a zinc finger protein: a database-guided approach. Proteins 1992;13:272.
48. Choo Y, Klug A. Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. Proc Natl Acad Sci USA 1994;91:11163–11167.
49. Segal DJ, Dreier B, Beerli RR, Barbas CF, 3rd. Toward controlling gene expression at will:selection and design of zinc finger domains recognizing each of the 5′-GNN-3′ DNA target sequences. Proc Natl Acad Sci USA 1999;96:2758–2763.
50. Bulyk ML, Huang X, Choo Y, Church GM. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proc Natl Acad Sci USA 2001;98:7158–7163.