Pacific Symposium on Biocomputing 6:115-126 (2001)

SAMIE: STATISTICAL ALGORITHM FOR MODELING INTERACTION ENERGIES

P.V. BENOS
Dept. of Genetics, Campus Box 8232,
Medical School, Washington University,
4566 Scott Ave., St. Louis, MO 63110, U.S.A.
e-mail: benos@genetics.wustl.edu

A.S. LAPEDES
Theoretical Division,
Los Alamos National Laboratories,
Los Alamos, NM 87545, U.S.A.
e-mail: asl@lanl.gov

D.S. FIELDS
MCB Biology, Univ. of Colorado,
Boulder, CO 80309, U.S.A.

G.D. STORMO
Dept. of Genetics, Campus Box 8232,
Medical School, Washington University,
4566 Scott Ave., St. Louis, MO 63110, U.S.A.
e-mail: stormo@genetics.wustl.edu

We are investigating the rules that govern protein-DNA interactions, using a statistical mechanics based formalism that is related to the Boltzmann Machine of the neural net literature. Our approach is data-driven, in which probabilistic algorithms are used to model protein-DNA interactions, given SELEX and/or phage data as input. In the current report, we trained the network using SELEX data, under the "one-to-one" model of interactions (i.e. one amino acid contacts one base). The trained network was able to successfully identify the wild-type binding sites of EGR and MIG protein families. The predictions using our method are the same or better than that of methods existing in the literature. However our methodology offers the potential to capitalise in quantitative detail, as well as to be used to explore more general model of interactions, given availability of data.

1 Introduction

Unraveling the general rules behind the recognition of specific DNA sequences by particular proteins has become a great challenge in computational biology. Many important biological processes depend on such accurate identification: DNA replication, methylation, and cell defense are among them. However, the most extensively studied such process is gene transcription, which is one of the

principal mechanisms of gene regulation. It is very important for the response of single celled organisms to environmental changes and essential for proper growth and development in multicellular organisms. It is mainly controlled by proteins that bind particular DNA target sequences, typically within the vicinity of the promoter region of the gene. These proteins affect the rate of transcription either positively (activators) or negatively (repressors), often through the action of complexes involving additional proteins.

Although the target sequences are relatively short, and therefore, abundant in the genome, DNA binding proteins are able to recognise them with high specificity. Much has been learned about these interactions in the past two decades [18]. There are now many protein-DNA complexes whose structures have been determined by X-ray crystallography. In addition, new techniques have allowed the determination of the preferred binding sites for many different proteins, wild type and mutants. Selection techniques, both *in vivo* and *in vitro*, have been applied to obtain either high affinity binding sites for particular proteins or high affinity proteins for particular DNA sites (e.g. [2,3]).

1.1 The search for a "Protein-DNA Recognition Code"

The search for the "Protein-DNA Recognition Code" has been a long time pursuit. Such a code would allow one to predict the binding site for a protein by knowing its sequence (and inferring its structure by homology to other proteins of that family) or *vice versa*. Moreover, having a recognition code would allow for the design of proteins that bind particular sequences and would open new horizons in manipulating gene expression.

In 1976, Seeman *et al*[22] proposed a rational protein-DNA recognition code based on the surface features of the amino acids and the bases. However, by the time a few protein-DNA complexes had been crystallized it was clear that such a simple code was not realistic. A 1988 *Nature* paper entitled "Protein-DNA interaction. No code for recognition" [16] showed that there is no simple, deterministic recognition code, as in the genetic code. But even the genetic code is deterministic in only one direction: given a codon we know with certainty the corresponding amino acid, but given the amino acid we only know the frequencies of various codons. The protein-DNA recognition code is clearly probabilistic in both directions. There are clear preferences for given amino acids to interact with particular base pairs and *vice versa* [17,5,12,15,24].

We are using a data-driven approach to incorporate these preferences into a well defined probabilistic code. Previous attempts for determining such a "recognition code" include the exploitation of DNA-protein co-crystal structural data [15] as well as the development of a *qualitative* model [4,24]. However,

the first of these approaches is limited by the small size of the data set (53 examples); whereas the second one is bound to the "one-to-one" model of interactions and the binary representation of the data (see also Sec. 4.2).

1.2 EGR protein family: a brief overview

In recent years a number of protein families has been used for the study of protein-DNA interactions, but the best studied so far is the family of Early Growth Response factors (EGR). EGR genes are early-response genes, first identified in mammals[7]. In later years they were also cloned (fully or partially) from a variety of other organisms, including zebrafish and *Xenopus laevis*. All of those genes contain three highly conserved zinc finger regions, a domain common to many eukaryotic transcription factors. The structure of the three Zn-fingers of the protein bound to its consensus DNA sequence was initially solved crystallographically at 2.1 $\mathring{A}$[19] and consequently refined to 1.6 $\mathring{A}$[6]. The target site is now believed to be 10 bp long and each finger contacts 4 of these bases (with one base overlap in the target site of each finger)[6]. The topology of the molecules in the solved crystal structure showed that each of four "critical" amino acids in every finger could contact a base on the target site (Figure 1). It was also found that the three fingers bind the DNA in a modular fashion, independently of each other[2,20].

2 Data

2.1 Data Sources: SELEX and Phage Display Experiments

There are three types of interaction data currently available in the literature. The first is SELEX data, where a particular protein is used to fish out oligonucleotide target sequences from a randomised pool[2]. The second type is data derived from phage-display experiments, where the DNA target site is fixed and the protein is randomised[3]. Both kinds of experiments allow one to select the variable part (DNA target sites or binding proteins) that has high relative affinity to the fixed part (proteins or DNA respectively). However, usually several different sequences are obtained from these experiments and we do not know their relative affinities, or even if these are the only high affinity sites. We can only infer that the observed sites are among the highest affinities sites for that protein/DNA. A third type of data comes from experiments where both the protein (wild-type or mutated) and the DNA target are fixed. These data are not informative for our probabilistic approach.

We collected from the literature 876 examples of DNA bound by variants of EGR proteins. 367 of these resulted from SELEX and 274 from phage-display

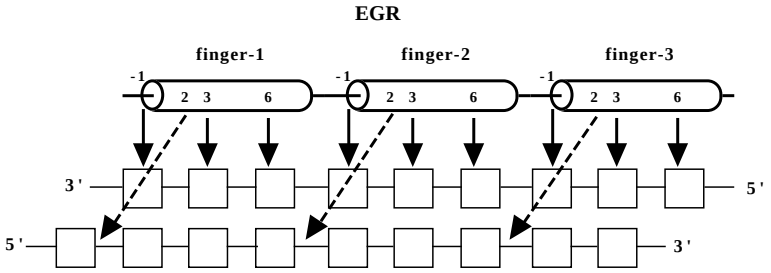

Figure 1: **"One-to-one" model of interaction.** DNA binding model for the EGR protein family, according to the crystallographic studies. EGR proteins have three zinc finger domains, each of which contacts four bases in an antiparallel fashion. There is an one base overlap in the target sequence between any two adjacent fingers. The numbering of the amino acids is with respect to the beginning of the alpha-helix. Amino acids -1, 3 and 6 contact bases at positions 3, 2 and 1 respectively; whereas amino acid 2 contacts the complementary base at position 4 (overlapping base).

experiments. In 235 cases both the protein and the DNA target were fixed.

In the present report, we focus on SELEX data from studies on EGR-derived proteins. According to "one-to-one" model of interaction[6], amino acids at positions -1, 3 and 6 (with respect to the beginning of the α-helix) contact bases at positions 3, 2 and 1 respectively; whereas amino acid at position 2 contacts the complementary base at position 4 (overlapping base between two adjacent fingers, as shown in Figure 1).

Zn-fingers of this type (C2H2) are believed to function in a modular fashion, independently of each other (except the overlapping base)[2,20]. We decided to focus our analysis on the interactions of a single finger. We created a dataset of 1,101 training vectors, by pooling together all single fingers from the 367 SELEX experiments. Since our approach is statistical by nature, we discarded the 426 of those examples, that both the DNA and the protein was fixed in all positions (see example, below). Thus, we ended up with 675 single finger vectors, which constituted our training set. In this set 115 different "proteins" (with respect to the four "critical" amino acids) had "selected" a total of 52 different tetranucleotide targets (out of all possible 256).

As an example of the training set construction, consider the following SELEX result (capital and small letter denotes randomisation and fixation of the corresponding base and $f1$, $f2$ and $f3$ represent the sequences of the three fingers; "critical" positions -1, 2, 3 and 6 of each finger appear in bold).

$$5'gcgGTGgcgt3' - - - f1s\mathbf{rs}\mathbf{de}lt\mathbf{r}hir - f2s\mathbf{r}v\mathbf{da}le\mathbf{a}hrr - f3a\mathbf{r}s\mathbf{de}rk\mathbf{r}htk$$

From this particular experiment, two training vectors were obtained: '$gcgG$ –

rder' (finger-3) and '$GTGg - rdaa$' (finger-2). Finger-1 ('$gcgt - rder$') was excluded from the training set, since all four bases of its DNA target are fixed. Finally, under the "one-to-one" model (Figure 1), only a small subset of all possible interactions between bases and amino acids was considered. For example, for the vector '$GTGg - rdaa$', the allowed interactions were: first G to second Ala, T to first Ala and second G to Arg.

2.2 Data representation

The data can be represented as two sparsely encoded vectors (${}_xN$ and ${}_yA$), which consist of the binary nature of the representation of the target DNA and the "critical" amino acids respectively. For the binary representation of the four bases we use the following vectors:

$$A = N^1 = (1000), C = N^2 = (0100), G = N^3 = (0010), T = N^4 = (0001)$$

Using this notation, a four base long string at positions $i = 1$ to 4 is represented as a string of *vectors* at the four positions, N_i^α. Similarly, there is another set of twenty such vectors for the representation of the amino acids.

The amino acids are assumed to interact with the DNA in a mode that is provided by a "contact matrix", C. Matrix C consists of binary values: if base at position i contributes to the affinity of interaction by contacting amino acid at position j then $C_{ij} = 1$, otherwise it is 0.

3 The Algorithm

Problem: Given a fixed and a variable counterpart (e.g. protein[a] and DNA respectively), find out which of the combinations of the variable counterpart have high specificity to the fixed one.

3.1 The model

We assume that an effective potential, E, of the following form exists, which describes the binding of protein to DNA:

$$E({}_xN, {}_yA) = \sum_{ij\alpha\beta} C_{ij} T_{ij}^{\alpha\beta} \; {}_xN_i^\alpha \; {}_yA_j^\beta + \sum_{i\alpha} H_i^\alpha \; {}_xN_i^\alpha + \sum_{j\beta} J_j^\beta \; {}_yA_j^\beta \qquad (1)$$

[a] The term "protein", here, simply refers to those amino acids that contact the DNA (or we assume they do so).

Here $E({}_xN,{}_yA)$ is the **effective energy** of binding of the nucleotide sequence ${}_xN$ to amino acid sequence ${}_yA$. On the right side of the equation, ${}_xN$ and ${}_yA$ are decomposed into the individual residues (the αs and βs) and positions (the is and js) that make up the sequences, so as to specifically include the additive contributions from each possible base-amino acid contact. $T_{ij}^{\alpha\beta}$ represents an additive contribution to the energy resulting from a nucleotide, α, to amino acid contact, β, at position pair (i,j) (see Figure 2B). For completeness we include H_i^α and J_j^β, which represent possible position dependent contributions to the energy, independent of interaction (for bases and amino acids, respectively). C_{ij} is the "contact matrix" we referred to before. By modifying the C_{ij} matrix we can explore various models that allow different subsets of interactions.

It is easy to extend this form to include more complicated interactions. For example, non-additive di-nucleotide interactions with residues can be represented in an extended form of T_{ij} matrix. However, in the present study we focus on the above form, where we assume that each nucleotide-residue contact makes an additive contribution, according to the "one-to-one" model.

3.2 Specificity

Consideration of the binding of protein to DNA occurring within the cell [1,9] prompts the form of the algorithms we use to analyse the available sequence data. Previous work [9], employed the idea of *specificity* in sequence analysis of protein-DNA binding interactions. Assume that nucleotide sequences exist with some reference probability, $P_{ref}({}_xN)$. Then, following Berg and von Hippel [1], we write the probability that a protein will bind to a typical sequence in the set of sequences described by the reference distribution as:

$$P({}_xN|{}_yA) = \frac{P_{ref}({}_xN)exp(-E({}_xN,{}_yA))}{\sum_{zN} P_{ref}({}_zN)exp(-E({}_zN,{}_yA))} \tag{2}$$

where E is an assumed effective potential of interaction, taken from Equation 1; and $\sum_{zN}$ denotes a sum over all possible nucleotide sequences. The fraction of time that the protein will be bound to *one* of the nucleotide sequences (there can be multiple sequence copies with varying multiplicities for each sequence) is the **specificity**, denoted by K:

$$K({}_xN|{}_yA) = \frac{P({}_xN|{}_yA)}{P_{ref}({}_xN)} \tag{3}$$

and the **reference probability**, $P_{ref}({}_xN)$ will be that determined by the nucleotide frequencies for this experiment.

Given a set of SELEX generated sequence data the overall specificity is just the product over the data items of each individual specificity and thus we can write:

$$logK^{total} = \sum_{xy} logK({}_xN|_yA) \tag{4}$$

The $\sum_{xy}$ is over all of the base-amino acid sequence combinations in the database. Similar expressions can be written for the specificity involved in a phage display experiment.

3.3 Maximising Likelihood

Assuming that the sequences were selected for optimal specificity, the parameters T,H,J can be determined by maximising the log probability (or the log specificity) for all the data as a function of these parameters. This process is called *parameter fitting* .

In our case, the model was trained by adjusting the weights according to the **steepest ascents** procedure, with which any objective function can be maximized by iteratively incrementing each free parameter by an amount proportional to its gradient with respect to the parameters. This process has a very simple interpretation for our log specificity objective function. It can be written as the sum of the averages as calculated by frequency counting in the given data set and the (negative of) the expectation as computed within the distribution. *Hence the steepest ascent process will reach a fixed point (i.e. zero gradient) when the expectations as calculated within the distribution match the frequencies as calculated within the given data set.* This intuitive result is basically the *Boltzmann machine* algorithm for neural network training [8], an observation which results in some additional insights into the algorithm we propose, but won't detail here.

4 Results and Discussion

We used four different data sets to test our model: one was the training set itself ("self-test"), two others were obtained from more recent publications on the EGR protein family [23,24] and the fourth consisted of data on MIG proteins [14]. MIG are also Zn-finger proteins of the same type (C2H2), although unrelated to EGR outside the Zn-finger domains.

4.1 Self-test

We examined how well the model predicts its own training set. If the model is internally consistent, then the combinations used on its training phase ought, in general, to have high probability rank. Of course, we don't expect that all of the training set combinations will be ranked the highest. In a SELEX experiment, many DNA targets may be selected, but they cannot all be predicted to be of the highest rank, although they should be near the top. Moreover, the stochastic nature of the experimental procedure will not always select the sequences with the highest probability. The ones selected, though, should have probability near to the highest.

Thus, for each of the 115 fingers of the training set (in fact: tetra-peptides; see also Sect. 2.1), we ranked all possible 256 tetra-nucleotide targets according to our model's predicted specificity. We found that 80% of the DNA targets of the training set rank in the top 1% of the list of all possible targets (of their associated finger). This is a reasonably good result which we expect to become better when phage display data is included during training and when larger sets become available for training.

4.2 Predicting SELEX binding sites on EGR-derived proteins

Wolfe *et al.* [24] performed a number of SELEX experiments using EGR-derived proteins. These proteins were originally optimised to bind DNA target sites normally recognised by NRE, TATA, and p53 proteins (named NRE_{ZF}, $TATA_{ZF}$ and $p53_{ZF}$ respectively). The data from these SELEX experiments were not included in our training data set.

We compared the predictions of the model for the three fingers of NRE_{ZF}, $TATA_{ZF}$ and $p53_{ZF}$ proteins with the predictions of the qualitative model Wolfe *et al.* proposed in their paper. This qualitative model can be viewed as another form of our quantitative, but with binary values and considering the "one-to-one" model of interaction only (Figure 2A). We found that in general, both models agree as to the preferred sites for each protein and usually the predictions match the observations. In a couple of cases, our model made the same predictions as theirs and neither matched the SELEX data (see Table 1). These examples probably indicate a limitation of the simple additive model that both employ and are worth exploring in more detail experimentally. In two cases their model made no prediction ("N") whereas ours made one that is consistent with the data. There is no case where the qualitative model does better than ours.

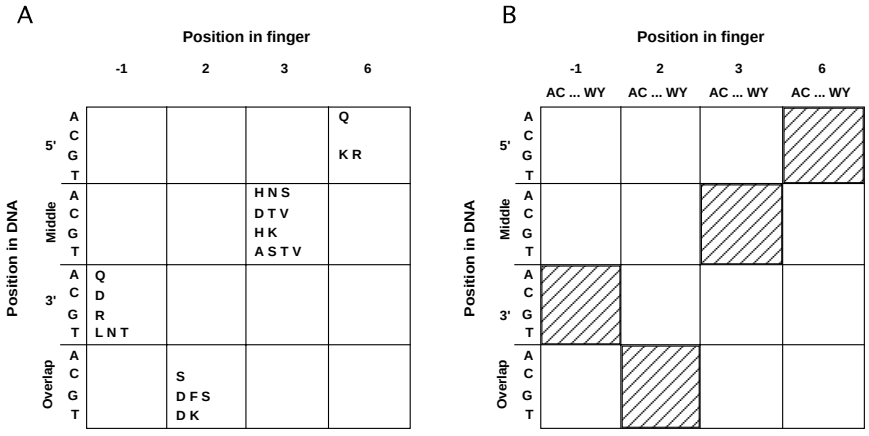

Figure 2: Schematic view of the two models: (**A**) qualitative and (**B**) quantitative. The boxed areas of the quantitative model correspond to the "one-to-one" model of interaction. The qualitative model could also be viewed as a quantitative with binary values only.

4.3 Quantitative Predictions on Relative Affinity

We compared our predictions to some quantitative data that were reported in [23]. Among other experiments, they measured the change in affinity between two different binding sites for a number of protein variants (summarised on Table 1, p. 2762 of their paper). They report 9 such differences; but because our training data did not include an example of Lys in amino acid position 3, we can only compare our predictions to 8 of their measurements. Their observed differences in affinity range from essentially no change to over 160-fold. If we calculate the expected change in affinity for those pairs, based on our model, the range is not as large as the measured one; but there is a strong correlation between the two (r=0.80). Moreover, in every case a change resulted in large differences in affinity, i.e. 30-160 fold, the predictions had large differences too. Similarly, changes that had minor effects on affinity, less than 5 fold, were predicted to have little or no effect.

4.4 Predicting S. cerevisiae MIG binding sites

S. cerevisiae has no EGR homologue. However, two yeast transcription factors (MIG1 and MIG2) contain Zn-fingers that belong to the same class as the ones of EGR proteins (C2H2). The Zn-finger region of the two protein families is highly similar (43.5% a.a. identity), although there is no conserva-

Protein	AA	DNA pos.	Qualitative	Quantitative	SELEX
$p53_{ZF}$	$fing2 : R^6$	1	G	G	A
NRE_{ZF}	$fing2 : D^{-1}$	3	C	C	n
$p53_{ZF}$	$fing1 : E^6$	1	n	AC	C
NRE_{ZF}	$fing3 : A^6$	1	n	AC	Ac

Table 1: Summary of the major discrepancies between the two prediction models, qualitative and quantitative, with respect to the SELEX results, reported at Wolfe *et al.* The first two rows show the cases that both qualitative and quantitative models made the same prediction, which didn't match the SELEX results. The other two rows show the cases where the qualitative model could not make any prediction. In both these cases our quantitative model predict A or C for the first DNA position; in the first case the observed consensus is C and in the second the consensus is A with C being the second most common base.

tion whatsoever in the rest of the proteins. Moreover, MIG proteins contain two Zn-fingers at the N-terminus, whereas EGR proteins contain three fingers at the C-terminus. Since they are, essentially, different proteins with a common function (they all bind DNA and presumably, in the same fashion), they constitute good candidates to test the prediction potentials of our method.

Mark Johnston and his colleagues performed SELEX experiments on the wild type MIG proteins[14]. In these experiments, they used MIG1 and MIG2 to select oligonucleotide targets from a randomised pool, biased for the SUC2-A MIG1 target site. The SUC2-A site is $5' - ATAAAAAT\underline{GCGGGGA}A - 3'$ and the oligonucleotide pool was biased to contain, in each position, 79% percent of the "wild-type" nucleotide and 7% of each of the other three. Assuming that MIG proteins bind the DNA in a fashion similar to EGR, our model predicts that their preferred binding site should be $5' - GCGGGGG - 3'$, which is exactly the result of their SELEX experiment. In addition, the model predicts that the most variable base is the last one, which is also confirmed by the SELEX result. In the second position, C is predicted to be the second most variable base, which agrees with the observed data. However, not everything about the model is consistent with the data. For example, the model predicts that the first position should be conserved the most, whereas *in vivo* studies have shown that the protein is more tolerant of changes there.

5 Concluding remarks

In this paper, we presented a probabilistic method that addresses the problem of DNA recognition by particular proteins. We developed a simple proba-

bilistic, data driven algorithm, which can "learn" from SELEX and/or phage display data, by optimising an objective function that is related to the specificity of the protein to DNA. Previous attempts to determine a set of "recognition rules" of the protein-DNA interactions are limited by the number of data used (53 co-crystal structures, as in [15]) or the nature of the model itself ("qualitative" model, as in [24]). Our method assumes additivity of the interactions between bases and amino acids, although it can be easily expanded to include non-additive interactions.

As a first approach, we used SELEX data from the EGR protein family, collected from the literature, for the training according to "one-to-one" model of interactions. We tested our model, by comparing its predictions to the training set itself and to data that were not included in it. In all cases, our method performed the same or better than the currently available qualitative model. Moreover, it can be easily expanded to explore more complex interaction patterns. We also used our model to make some quantitative predictions and compare them with affinity measurements from Segal *et al* [23]. We found that our predictions agreed with the observations (correlation factor r=0.80). Such quantitative predictions are not possible with any of the models available in the literature, and are one of the primary advantages of our approach. Using more data and joint training we expect to be able to make even more accurate predictions. Finally, we explored the potentials of predicting binding sites for yeast MIG1 and MIG2, two proteins with Zn-finger domain structure, similar to EGR. Our results confirm the outcome of the SELEX experiment for these proteins and thus indicates that they bind DNA in an antiparallel fashion, analogous to EGR.

The main limitation of our method, currently, appears to be the availability of data. We are planning to update our database and expand the training to include a combined SELEX and phage display data set. In addition, we are going to collect more data on C2H2 Zn-finger proteins and use them for a better training. However, there are more aspects to be addressed. Does the "additivity rule" hold? Are there interactions, other than the ones indicated in the crystal structure, that contribute significantly to specificity? We believe that the approach we presented here is on the right direction for a better understanding of the rules that govern the highly specific recognition of the DNA by proteins.

Acknowledgments

PVB wishes to thank Elena Rivas for useful discussion. The hospitality of the Santa Fe Institute, where part of this research was performed, is also

gratefully acknowledged. This work was supported by NIH Grant HG00249 to GDS. ASL's research was supported by the Department of Energy under contract W-7405-ENG-36.

References

1. Berg OG and von Hippel PH,*J Mol Biol* **193**, 723 (1987)
2. Choo Y and Klug A,*Proc Natl Acad Sci USA* **91**, 11163 (1994a)
3. Choo Y and Klug A,*Proc Natl Acad Sci USA* **91**, 11168 (1994b)
4. Choo Y, Klug A,*Curr Opin Biotechnol* **6**, 431 (1995)
5. Desjarlais JR, Berg JM,*Proc Natl Acad Sci USA* **89**, 7345 (1992b)
6. Elrod-Erickson M, Rould MA, Nekludova L, Pabo, CO,*Current Biology* **4**, 1171 (1996)
7. Gashler A and Sukhatme VP,*Prog Nucleic Acid Res* **50**, 191 (1995)
8. Hertz J, Krogh A and Palmer RG, "Introduction to the theory of neural computation". (Addison-Wesley Pub. Co., Redwood City, CA, 1991)
9. Heumann JM, Lapedes AS, Stormo GD,*ISMB* **2**, 188 (1994)
10. Isalan M and Choo Y,*J Mol Biol* **285**, 471 (2000)
11. Jamieson AC, Kim SH, Wells JA,*Biochemistry* **33**, 5689 (1994)
12. Jamieson AC, Wang H, Kim SH,*Proc Natl Acad Sci USA* **93**, 12834 (1996)
13. Klug A,*J Mol Biol* **293**, 215 (1999)
14. Lutfiyya LL, Iyer VR, DeRisi J, DeVit MJ, Brown PO, Johnston M,*Genetics* **150**, 1377 (1998)
15. Mandel-Gutfreund Y, Margalit H,*Nucleic Acids Res* **26**, 2306 (1998)
16. Matthews BW,*Nature* **335**, 294 (1988)
17. Nardelli J, Gibson TJ, Vesque C, Charnay P,*Nature* **349**, 175 (1991)
18. Pabo CO, Sauer RT,*Annu Rev Biochem* **61**, 1053 (1992)
19. Pavletich NP and Pabo CO,*Science* **252**, 809 (1991)
20. Pomerantz JL, Sharp PA, Pabo CO,*Science* **267**, 93 (1995)
21. Rebar EJ, Pabo CO,*Science* **263**, 671 (1994)
22. Seeman NC, Rosenberg JM, Rich A,*Proc Natl Acad Sci USA* **73**, 804 (1976)
23. Segal DJ, Drieter B, Beerli RR, Barbas 3rd CF,*Proc Natl Acad Sci USA* **96**, 2758 (1999)
24. Wolfe SA, Greisman HA, Ramm EI, Pabo CO,*J Mol Biol* **285**, 1917 (1999)
25. Wu H, Yang W-P, Barbas 3rd CF,*Proc Natl Acad Sci USA* **92**, 344 (1995)