Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 11099–11103, November 1994
Biochemistry

Length-encoded multiplex binding site determination: Application to zinc finger proteins

(DNA binding proteins/combinatorial libraries/molecular recognition)

JOHN R. DESJARLAIS AND JEREMY M. BERG*

Thomas C. Jenkins Department of Biophysics, The Johns Hopkins University, Baltimore, MD 21218; and Department of Biophysics and Biophysical Chemistry, The Johns Hopkins University School of Medicine, Baltimore, MD 21205

Communicated by Thomas D. Pollard, July 28, 1994

ABSTRACT **The screening of combinatorial libraries is becoming a powerful method for identifying or refining the structures of ligands for binding proteins, enzymes, and other receptors. We describe an oligonucleotide library search procedure in which the identity of each member is encoded in the length of oligonucleotides. This encoding scheme allows binding-site preferences to be evaluated via DNA length determination by denaturing gel electrophoresis. We have applied this method to determine the binding-site preferences for 18 Cys_2His_2 zinc finger domains as the central domain within a fixed context of flanking zinc fingers. An advantage of the method is that the relative affinities of all members of the library can be estimated in addition to simply determining the sequence of the optimal or consensus ligand. The zinc finger domain specificities determined will be useful for modular zinc finger protein design.**

Combinatorial libraries consist of sets of related compounds that can be searched for members that bind to a given receptor. Several schemes have been developed for producing encoded libraries in which a signal identifying each member is incorporated during library preparation. Signal incorporation has been achieved by positional encoding where each member is attached to a fixed location (1, 2) and by chemical methods in which beads with single ligands also include tags that can be read and decoded to reveal the identity of the ligand (3–5). For nucleic acid ligands, most libraries are screened by isolating single members and determining their sequences (6–9). We demonstrate a different scheme here in which sequence information is encoded in the lengths of DNA fragments via a multiplexing procedure. The length-encoding allows rapid readout via denaturing gel electrophoresis. A major advantage of our approach is that the properties of the entire library are simultaneously examined. This allows both the determination of the optimal ligands from the library and also quantitative evaluation of the level of discrimination at each monomer site. The determination of relative affinities is aided by the use of competition experiments with an equimolar mixture of all potential binding sites (10, 11).

We have applied this method to a series of designed zinc finger proteins. Zinc finger proteins of the Cys_2His_2 class typically contain tandem arrays of 28–30 amino acid domains (12, 13). On the basis of mutagenesis studies (14) and the crystal structure of the three zinc finger domains of Zif268 bound to DNA (15), it appeared that each zinc finger domain contacts a site consisting of three base pairs with the three-base subsites for each domain directly abutted. The apparent modular nature of these single-domain–three base subsite interactions suggested that novel DNA-binding proteins could be created by linking domains of known specificity. We have previously demonstrated that this is indeed possible through the use of previously determined zinc finger specificity rules and a consensus zinc finger framework (16). We have now constructed a zinc finger host–guest system in which the first and third domains of three-domain proteins are constant and the central domain is varied. We used the same design strategy developed previously (16), creating proteins which we term "QNR-XXX-RHR," [(Gln-Asn-Arg)-(Xaa-Xaa-Xaa)-(Arg-His-Arg)], where the letters refer to the single-letter code for the amino acids in the three positions most critical in determining DNA-binding specificity. The system is illustrated in Fig. 1. We constructed a total of 18 proteins with different combinations of amino acids in the X positions. Proteins of this form are expected to prefer DNA-binding sites of the form 5′-G(A or G)G-NNN-GA(T or A)-3′, since the preferred subsites of the domains termed QNR and RHR are expected to prefer 5′-GA(T or A) and G(A or G)G, respectively (17).

The specificities due to the central domains can be examined by testing the binding of these proteins to a collection of 64 potential sites composed of a variable three-base-pair site flanked by two fixed subsites recognized by the first and third zinc finger domains. The method that we have developed involves selection of the higher affinity sites from a pool of sequences of the form 5′-GAG-$N^1N^2N^3$-GAT-3′ in which the identity of the bases N^i has been encoded in the lengths of the fragments in the pool by synthesis as shown in Fig. 2.

MATERIALS AND METHODS

Construction of Genes and Protein Expression. A cassette for constructing the genes for the three zinc finger proteins was generated in the expression vector pG5 (18). This included coding regions for the leader and first and third zinc finger domains with *Xma* I and *Age* I restriction sites for insertion of fragments encoding the central zinc finger domains. Such fragments were synthesized individually or in sets and cloned into the cassette. The identity of all genes was confirmed by sequencing. Proteins were expressed and partially purified as described (16, 19).

Synthesis of Binding-Site Pools. The set of 12 pools of DNA probes was synthesized with the use of a MilliGen/Biosearch Cyclone DNA synthesizer. The two randomized positions in each pool were generated through the use of an equimolar mixture of all four phosphoramidites. The length-encoding of the different pools was accomplished in two steps. First, the identity of the fixed bases in each triplet (regardless of position) was encoded during oligonucleotide synthesis in the order A > C > G > T, where each set differed in length by one base from the previous set. Each pool included a palindromic sequence at the 3′ end so that the complementary strand could be generated by self-primed polymerase action. The sequences of the oligonucleotides synthesized are shown

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

*To whom reprint requests should be addressed.

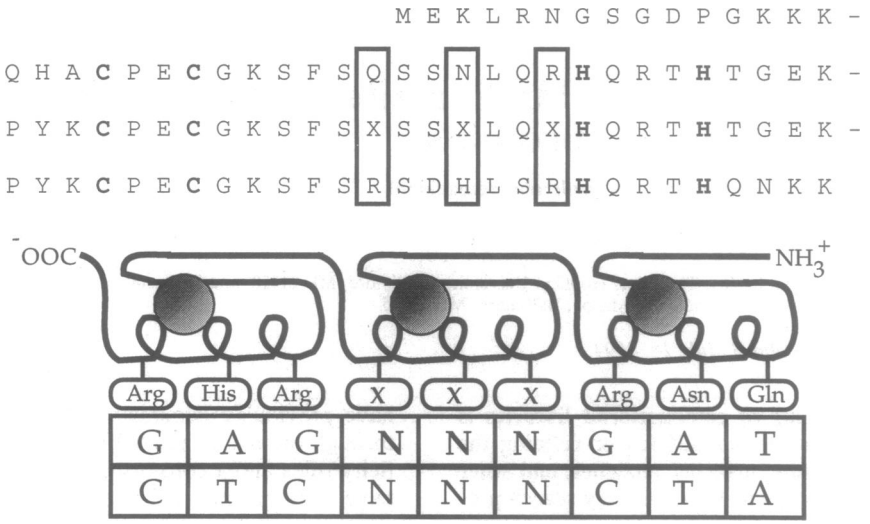

FIG. 1. (*Upper*) Amino acid sequences of the zinc finger proteins with fixed first and third domains and a central domain with variable contact residues. The sequences consist of the zinc finger consensus sequence CP-1 for each domain with altered contact residues and flanking sequences derived from Sp1 used previously. Note that the serine residue in position 15 in the central domain was changed to aspartic acid whenever arginine was used in position 13 in this domain. (*Lower*) Expected alignment of these proteins on their potential binding sites.

in Table 1. The double-stranded DNA fragments generated by self-primed polymerase reactions were cut with *Kpn* I and *Bam*HI and cloned into a slightly modified version of the pEMBL vector (20). Plasmid DNA was prepared from each pool and an equal amount of each was used for subsequent manipulations. Equal amounts of plasmid DNA for each set of four fixed bases were combined, cut with *Eco*RI, and end-labeled with [α-^{32}P]dATP. The remaining length-encoding was performed by cleaving each of these three mixtures of pools with different restriction enzymes in the polylinker of the vector. The remaining cleavage was done with *Bgl* II for (A, C, G, or T)NN sites, *Xho* I for N(A, C, G, or T,)N sites, and *Hin*dIII for NN(A, C, G, or T) sites. The restriction fragments were agarose gel-purified, extracted with β-agarase (New England Biolabs), and combined.

DNA Binding-Site Determinations. Mobility-shift selection from the length-encoded pools was performed on 1.8% Seaplaque (FMC) agarose gels. Approximately 10 ng of the labeled DNA pool were combined with protein in 35 mM Tris chloride, pH 8.0/60 mM KCl/90 μM $ZnCl_2$/3 mM dithiothreitol/300 μg of bovine serum albumin per ml/20 μg of poly (dI-dC) per ml/10% (vol/vol) glycerol buffer. After 15 min at room temperature, the samples were electrophoresed for 3.5 hr at 100 mA, and the gel was dried. The concentration of each protein was empirically adjusted so that 1–5% of the total pool was shifted. After visualization by autoradiography, the protein–DNA complex bands were excised, rehydrated, and digested with β-agarase. The DNA was precipitated with ethanol, resuspended in Sequenase loading buffer (United States Biochemical), and loaded onto an 8% polyacrylamide sequencing gel. A portion of the unselected pool was also loaded as a standard. The resulting patterns were visualized by autoradiography and quantitated with the use of a PhosphorImager (Molecular Dynamics). Band intensities were evaluated either by integration over the area of each band or by measuring peak heights on a trace derived from vertical integration across a lane containing the bands. Relative fractional saturation values for each pool were calculated by correcting the intensity of each selected band by the intensity of the corresponding band from the unselected standard. Normalized base propensities were calculated by dividing each relative fractional saturation value by the sum of the relative fractional saturation values at each position.

Determination of Individual Binding-Site Relative Free Energies. Thirteen individual binding sites were obtained by

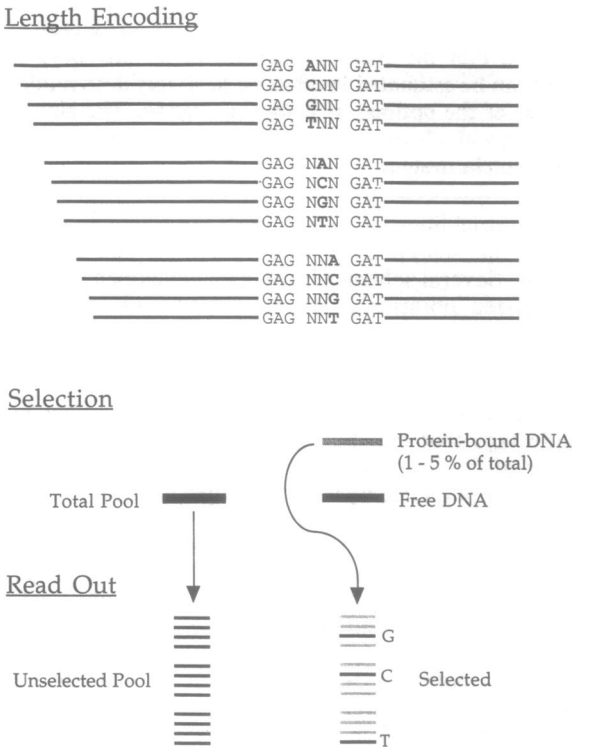

FIG. 2. Schematic view of the length-encoded multiplex binding-site determination system. Twelve pools of DNA fragments are prepared, each of which is a mixture of 16 different sequences. In the first pool, the identity of the first base on the variable control triplet is fixed as an A, whereas the other two sites with the triplet contain equal amounts of all four bases. The second pool has the identity of the first base fixed as C and the fragments in this pool are one base pair shorter than those in the first pool, the length having been set during pool synthesis. This is repeated as shown for the 12 pools. After preparation, these pools are mixed and end-labeled. Those members of the pool that are most tightly bound by a protein under study are selected by using a mobility-shift gel to separate bound from free DNA. The protein concentration is adjusted so that 1–5% of the total pool is shifted. After selection, the DNA from the protein–DNA complex is examined on a sequencing gel followed by autoradiography. The length-encoding allows the sequences of the preferred binding sites to be deduced directly from the pattern of bands produced as shown for a hypothetical protein that prefers to bind to 5′-GAG GCT GAT-3′.

Table 1. Synthesized oligonucleotides

Pools	Oligonucleotide sequence
ANN, NAN, NNA	5′-CTCTGGATCCACCA GAG NNN GAT TGGTACCAAT-3′
CNN, NCN, NNC	5′-CTCTGGATCCCCA GAG NNN GAT TGGTACCAAT-3′
GNN, NGN, NNG	5′-CTCTGGATCCCA GAG NNN GAT TGGTACCAAT-3′
TNN, NTN, NNT	5′-CTCTGGATCCA GAG NNN GAT TGGTACCAAT-3′

subcloning from the length-encoded pool. The relative dissociation constants for the complexes between these sites and the protein termed "QNR-QDR-RHR" (in which QDR-Gln-Asp-Arg) were determined by using gel mobility-shift assays under conditions identical to those used for the gel mobility-shift selection experiments.

RESULTS AND DISCUSSION

Specificity Due to XXX = QDR in the Central Domain. As an initial trial of our method, the DNA binding properties of the protein containing the residues QDR (Gln-13, Asp-16, and Arg-19) in the central XXX positions were examined. Domains containing these contact residues have been previously examined both as the central domain within the Sp1 framework sequence (19, 21) and in the consensus sequence framework (16). The pattern from the length-encoded multiplex assay for this protein is shown in Fig. 3 *Upper*. These data are shown in the form of a histogram of normalized base propensities at each of the three positions in Fig. 3 *Lower*.

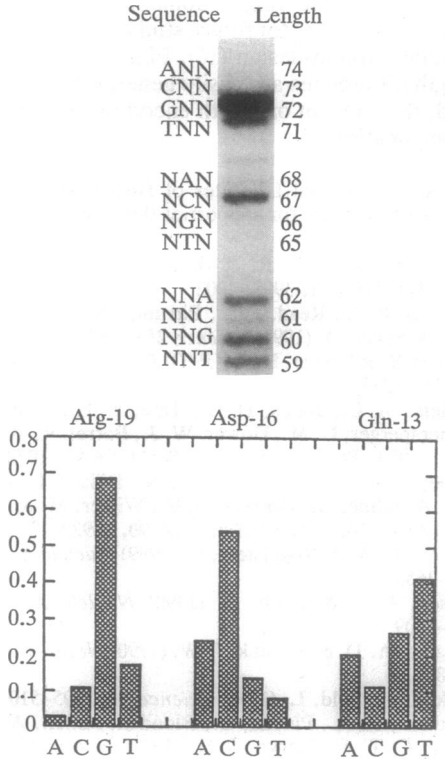

FIG. 3. Binding-site determination for the protein with XXX = QDR. (*Upper*) Length-encoded multiplex profile for this protein. The pattern is read from the top to the bottom in the 5′ to 3′ direction. Thus, the top four bands correspond to the preferences at the underlined position 5′-GAG-NNN-GAT-3′. The third band within this set is prominent, indicating a strong preference for G in this position. The middle set of four bands reveals the preferences in the central position. The intensity of the second band indicates a clear preference for C. The bottom set of four bands corresponds to the position 5′-GAG-NNN-GAT-3′. Much less discrimination is observed for this position. (*Lower*) Histogram for the data in (*Upper*) corrected for the intensities of the bands from the unselected pool and normalized. These correspond to relative base propensities.

Examination of these data reveals a strong selectivity for G determined by Arg-19, a slightly weaker preference for C determined by Asp-16, and a more modest preference for T with significant binding of A and G determined by Gln-13. These results exactly parallel those obtained in the other contexts noted above (16, 19, 21).

An important feature of the method is its potential for evaluating the relative affinities of each protein for all 64 possible binding sites. This depends of the extent of additivity of the relative free energy contributions from each base pair in each position of the $N^1N^2N^3$ subsites. To examine these effects, we determined the relative dissociation constants (and, hence, the relative binding free energies) for a set of 13 individual binding sites isolated from the length-encoded pools by individual gel mobility-shift assays. The probe containing the sequence GAG-GCT-GAT was found to be bound most tightly. The sequences of the central three base pairs for the 13 binding sites and their binding free energies (in kcal/mol) relative to the GCT site are as follows: GCT, 0; GCG, 0.4; TCA, 1.5; GCC, 1.7; GAG, 2.0; TTT 2.3; GAA, 2.7; ACG, 2.7; CCA, 3.0; ACT, 3.2; CGT, 3.5; TTC, 3.5; CGA, 3.9. We used these data to estimate specificity free energies, assuming independence between the three base positions. This was accomplished by fitting the relative free energies for each site to the expression $\Delta G_{calc} = \Delta\Delta GN^1 + \Delta\Delta GN^2 + \Delta\Delta GN^3$, where the variables are the $\Delta\Delta GN^i$ values, the relative free energy contributions for the base N at position i (relative to the most favored base at each position). These nine $\Delta\Delta GN^i$ values were adjusted to reproduce the relative affinity data for the 13 binding sites. A correlation coefficient of 0.98 was obtained between the observed and calculated ΔG values. These results were converted to normalized base propensities for each position i via the expression $f_X = e^{-(\Delta\Delta GX_i/kT)}/\Sigma[e^{-\Delta\Delta GN_i/kT)}]$ for X = A, C, G, and T with the summation over N^i = A, C, G, and T. Relative free energies were calculated from the multiplex profile distribution via the expression $\Delta\Delta GX^i = -kT\ln(f_XQ)$, where f_X is the normalized base propensity for base X in position i and $Q = 1/f_X(\max)$ in which $f_X(\max)$ is the normalized base propensity for the most favored base at

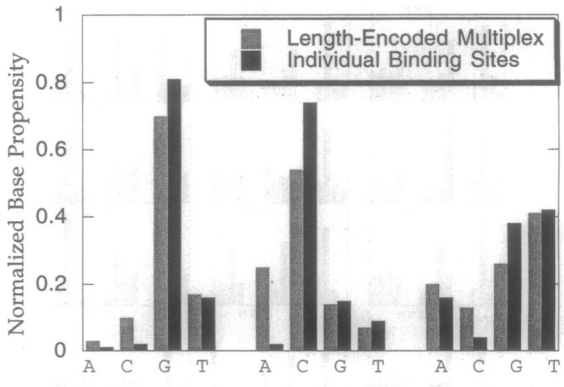

FIG. 4. Comparison of length-encoded multiplex data and those derived from individual binding-site free energies. Normalized base propensities obtained directly from the length-encoded multiplex binding assay are compared with values derived from the binding free energies for 13 individual sites by assuming independence between the three positions in the center of the binding site.

position *i*. This expression applies in the limit of low binding saturation. These relative free energies were compared with those derived from the individual binding experiments. A correlation coefficient of 0.83 was found. The normalized base propensities deduced from the individual sites and those from the length-encoded multiplex profile are directly compared in Fig. 4. Impressive agreement is achieved, suggesting that the individual site preferences are reasonably additive. The degree of discrimination at each position appears to be slightly underestimated by the multiplex profile method. This may be due, in part, to the effects of slight band overlap on the integration methods used to convert the experimental data to the normalized preferences. Overall, the level of agreement validates the use of the length-encoded multiplex method for the generation of the thermodynamic discrimination levels between optimal and suboptimal sites.

Examination of Additional Proteins. Seventeen additional proteins were examined by the length-encoded multiplex binding assay. A representative set of profiles is shown in Fig. 5 *Upper*, and the complete set of histograms is shown in Fig. 5 *Lower*. The results reveal a range of specificities. For example, comparison of the QDR profile discussed above with that for QNR shows that the specificity in the central position is changed from C to A, paralleling results observed on other contexts (17). In addition, however, the pattern in the 3′ position in the triplet also changed somewhat showing less preference for T and more for A. This effect was strongly enhanced for the QNN (Gln-Asn-Asn) protein, which showed a strong preference for GAA. The preference for G in the 5′ position is interesting in that the arginine residue often in this position and known to form two hydrogen bonds to G in this position in the Zif268 cocrystal structure (15) has been replaced with the much shorter asparagine residue with little change in specificity. The basis for this specificity remains to be determined. The RHE (Arg-His-Glu) protein shows a clear preference for TGG, suggesting discrimination for T by Glu-19 in this context. Finally, not all of the proteins show any clear specificity. For example, the DAR (Asp-Ala-Arg) protein shows very little discrimination in any position.

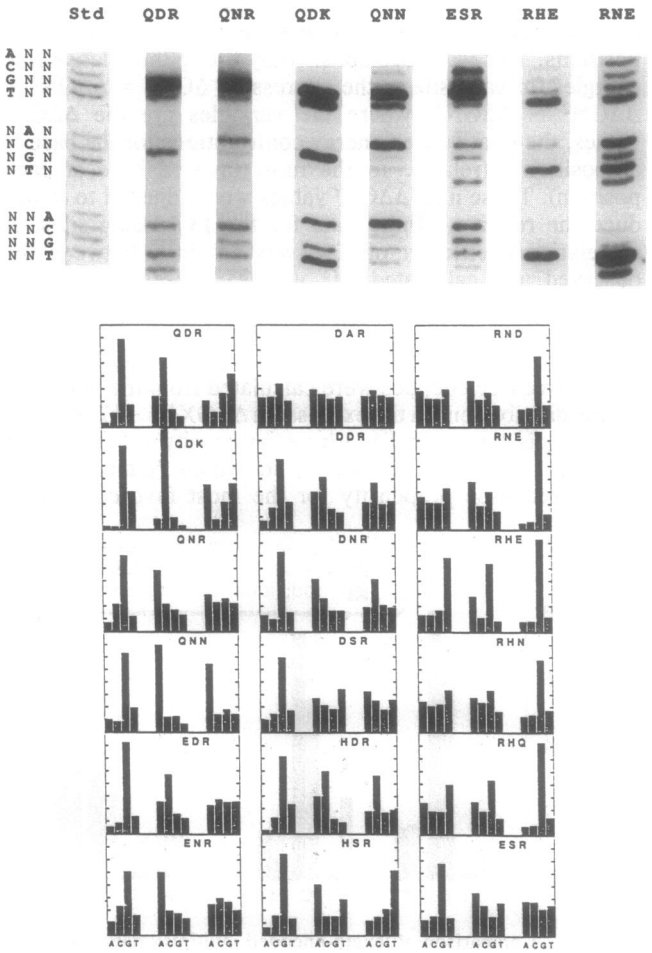

FIG. 5. Length-encoded multiplex profiles for 18 zinc finger domains. (*Upper*) Experimental profiles for a representative set of proteins. (*Lower*) Normalized base propensity histograms for the 18 proteins.

Conclusions. A variety of methods for identifying sites for DNA binding proteins have been developed (6, 7, 22). These generally involve multiple rounds of selection and amplification from a random or randomized collection of binding sites and are very powerful for determining the optimal or consensus site for a given protein. Because of their multiple round character, however, these methods can be labor intensive and the results are not readily interpretable in terms of thermodynamic preferences. Such thermodynamic information has been available only for a small number of DNA-binding proteins. One example is λ repressor for which binding free energies were determined for a natural binding site and for all possible single-point mutations by individual binding titrations monitored by filter binding (23). The length-encoding multiplex method described herein efficiently provides a similar level of information about both the optimal binding site and about the levels of discrimination at each position within the site. The data obtained has provided an expanded set of zinc finger domains with characterized binding site preferences for future studies of zinc finger-based DNA binding protein design (16). Finally, since oligonucleotide-length differences are easily generated, measured, and compared, the concept of length encoding should find utility in other applications.

We thank the National Institutes of Health and the Lucille P. Markey Charitable Trust for support of this work.

1. Geysen, H., Meloen, R. & Barteling, S. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 3998–4002.
2. Fodor, S. P. A., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T. & Solas, D. (1991) *Science* **251**, 767–773.
3. Brenner, S. & Lerner, R. A. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 5381–5383.
4. Needels, M. C., Jones, D. G., Tate, E. H., Heinkel, G. L., Kochersperger, L. M., Dower, W. J., Barret, R. W. & Gallop, M. A. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 10700–10704.
5. Ohlmeyer, M. H. J., Swanson, R. N., Dillard, L. W., Reader, J. C., Asouline, G., Kobayashi, R., Wigler, M. & Still, W. C. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 10922–10926.
6. Kinzler, K. W. & Vogelstein, B. (1989) *Nucleic Acids Res.* **17**, 3645–3653.
7. Thiesen, H.-J. & Bach, C. (1990) *Nucleic Acids Res.* **18**, 3203–3209.
8. Ellington, A. D. & Szostak, J. W. (1990) *Nature (London)* **346**, 818–822.
9. Tuerk, C. & Gold, L. (1990) *Science* **249**, 505–510.
10. Houghton, R. A., Pinella, C., Blondelle, S. E., Appel, J. R., Dooey, C. T. & Cuervo, J. H. (1991) *Nature (London)* **354**, 84–86.
11. Zimmermann, R. N., Kerr, J. M., Siani, M., Banville, S. C. & Santi, D. V. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 4505–4509.
12. Klug, A. & Rhodes, D. (1987) *Trends Biochem. Sci.* **12**, 464–470.
13. Berg, J. M. (1990) *Annu. Rev. Biophys. Biophys. Chem.* **19**, 405–421.
14. Nardelli, J., Gibson, T. J., Vasque, C. & Charnay, P. (1991) *Nature (London)* **349**, 175–179.
15. Pavletich, N. P. & Pabo, C. O. (1991) *Science* **252**, 809–817.
16. Desjarlais, J. R. & Berg, J. M. (1993) *Proc. Natl. Acad. Sci. USA* **90**, 2256–2260.
17. Desjarlais, J. R. & Berg, J. M. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 7345–7349.

18. Alexander, P., Fahnestock, S., Lee, T., Orban, J. & Bryan, P. (1992) *Biochemistry* **31**, 3597–3603.
19. Desjarlais, J. R. & Berg, J. M. (1992) *Proteins Struct. Funct. Genet.* **12**, 101–104.
20. Dente, L., Cesareni, G. & Cortese, R. (1983) *Nucleic Acids Res.* **19**, 1645–1654.
21. Desjarlais, J. R. & Berg, J. M. (1992) *Proteins Struct. Funct. Genet.* **13**, 272.
22. Woodring, W. E. & Funk, W. D. (1993) *Trends Biochem. Sci.* **18**, 77–80.
23. Sarai, A. & Takeda, Y. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 6513–6517.