16. Vinson, C. R., Sigler, P. B. & McKnight, S. L. *Science* **246**, 911-916 (1989).
17. O'Shea, E. K., Rutkowski, R. & Kim, P. S. *Science* **243**, 538-542 (1989).
18. Oas, T. G., McIntosh, L. P., O'Shea, E. K., Dalquist, F. W. & Kim, P. S. *Biochemistry* **29**, 2891-2894 (1990).
19. Ptashne, M. *Nature* **335**, 683-689 (1989).
20. Bohmann, D. & Tjian, R. *Cell* **59**, 709-717 (1989).
21. Brewley, M., Brovetto-Cruz, J. & Li, C. H. *Biochemistry* **8**, 4701-4708 (1969).
22. Abate, C., Luk, D., Gagne, E., Roeder, R. & Curran, T. *Mol. Cell biology* (in the press).

ACKNOWLEDGEMENTS. We thank P. Roepe for advice on fluorescence spectroscopy and M. Boublik for advice on circular dichroism.

Folding transition in the DNA-binding domain of GCN4 on specific binding to DNA

Michael A. Weiss*†, Thomas Ellenberger‡, C. Richard Wobbe*, Jonathan P. Lee*, Stephen C. Harrison‡ & Kevin Struhl*

* Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA
† Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts 02114, USA
‡ Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts, USA

PROTEIN–DNA recognition is often mediated by a small domain containing a recognizable structural motif, such as the helix–turn–helix[1] or the zinc-finger[2]. These motifs are compact structures that dock against the DNA double helix. Another DNA recognition motif, found in a highly conserved family of eukaryotic transcription factors including C/EPB, Fos, Jun and CREB, consists of a coiled-coil dimerization element—the leucine-zipper—and an adjoining basic region which mediates DNA binding[3]. Here we describe circular dichroism and ^{1}H-NMR spectroscopic studies of another family member, the yeast transcriptional activator GCN4[4,5]. The 58-residue DNA-binding domain of GCN4, GCN4-p, exhibits a concentration-dependent α-helical transition, in accord with previous studies of the dimerization properties of an isolated leucine-zipper peptide[6]. The GCN4-p dimer is ~70% helical at 25 °C, implying that the basic region adjacent to the leucine zipper is largely unstructured in the absence of DNA. Strikingly, addition of DNA containing a GCN4 binding site (AP-1 site) increases the α-helix content of GNC4-p to at least 95%. Thus, the basic region acquires substantial α-helical structure when it binds to DNA. A similar folding transition is observed on GCN4-p binding to the related ATF/CREB site, which contains an additional central base pair. The accommodation of DNA target sites of different lengths clearly requires some flexibility in the GCN4 binding domain, despite its high α-helix content. Our results indicate that the GCN4 basic region is significantly unfolded at 25 °C and that its folded, α-helical conformation is stabilized by binding to DNA.

The DNA-binding domain of the yeast transcription factor GCN4, located within the C-terminal 60 residues of the protein[7], can be divided into two subdomains. The C-terminal subdomain (32 residues) contains four leucine residues in a heptad repeat, the leucine zipper. Physicochemical studies of a synthetic leucine-zipper peptide[6,8] and 'zipper swap' experiments[9,10] support the proposal[11] that the leucine repeat forms a hydrophobic dimerization interface. The *N*-terminal subdomain (the basic region; 26 residues) is rich in lysine and arginine residues, many of which are conserved in this family of transcription factors. Domain swap experiments demonstrate that the basic region has the predominant role in making specific DNA contacts[12].

To further our understanding of the structural basis for DNA recognition by this family of transcription factors, we have examined by circular dichroism (CD) and ^{1}H-NMR spectroscopy, the structure of a peptide (GCN4-p; residues 225–281) that includes the basic region and leucine repeat of GCN4. GCN4-p was overexpressed in *Escherischia coli*, and its DNA-binding properties were analysed by gel-retardation assay (Fig. 1). GCN4-p binds to synthetic DNA containing a consensus AP-1 target site with an affinity ($K_{\text{apparent}} = 2\times10^{-8}$ M; Fig. 1, lanes 1–6) comparable to that of intact GCN4 (ref. 4). GCN4-p fails to bind to an unrelated DNA sequence (the λ O_L1 operator; Fig. 1, lanes 13 and 14). These results show that the DNA-binding properties of the GCN4-p peptide are very similar to those of the intact GCN4 protein.

At high protein concentrations (>100 μM) the far-ultraviolet CD spectrum of GCN4-p shows the characteristics of α-helical structure (Fig. 2*A*). The calculated α-helix content of GCN4-p under these conditions is 70–73%, or 40–43 residues in helical conformation, calculated on a mean residue ellipticity at 222 nm and 25 °C of $-22{,}700$ deg cm^2 dmol^{-1}. No increase in helix content is seen at higher protein concentrations. Progressive attenuation of the α-helix-associated absorption bands is observed with decreasing peptide concentration (Fig. 2*A*), indicating an unfolding transition that is presumably accompanied by dissociation of the dimer into unstructured monomers. Previous studies of a leucine-zipper peptide (residues 252–281[6]) indicated that it is dimeric throughout the concentration range studied here. We suggest that decreased stability of GCN4-p dimers relative to those of the leucine-zipper peptide may result from repulsion of the basic region subdomains.

The unfolding of GCN4-p with decreasing concentration is also seen by ^{1}H-NMR spectroscopy. The leucine-zipper subdomain contains two aromatic residues, Tyr 265 and His 266, whose resonances are well resolved and serve as intrinsic probes of GCN4-p dimerization (Fig. 2*B*). At high and low peptide concentrations (Fig. 2*B*; spectra *a* and *f*, respectively), the resonance linewidths ($1/T_2$) of these residues are consistent with the presence of dimeric and monomeric species. Higher-order oligomers would be expected to show broader resonances, which are not observed. At intermediate peptide concentrations, the pairs of aromatic NMR resonances corresponding to monomeric and dimeric species have relative intensities that vary with GCN4-p concentration (Fig. 2*B*, spectra *b*–*e*). By contrast, NMR resonances of basic region residues (for example, Thr 236) show no intensity change over this range of peptide concentration (data not shown). These results unambiguously assign the concentration-dependent increase in α-helix content of GCN4-p Fig. 2*A*) to the leucine-zipper region of the peptide.

Previous studies of the dimerization properties of an isolated leucine-zipper peptide showed that the orientation of α-helices in the dimer is parallel, and it was proposed that the leucine zipper is a coiled-coil structure[6]. Two-dimensional NMR studies of this peptide shows that it forms a continuous α-helix, with symmetrical dimerization contacts that are evident from the single resonance observed for each proton in the peptide[8]. This symmetry is consistent with the parallel association of peptide α helices in a coiled-coil arrangement. The dimer-related protons of GCN4-p also show a single resonance (for example, His 265 and His 265′; Fig. 2*Ba*), suggesting that the dimerization interface of the GCN4-p DNA binding domain is similar to that observed for the leucine-zipper peptide. The estimated lifetime of GCN4-p dimers at 25 °C is between 10 ms and 1 s, on the basis of the exchange properties of NMR resonances assigned to the leucine-zipper region (see Fig. 2). In conjunction with the modest dissociation constant of GCN4-p for DNA, these observations suggest that unfolding and reassembly of GCN4, and other transcription factors utilizing a coiled-coil dimerization interface, may occur without significant kinetic barriers, thereby facilitating subunit exchange. Such exchange is thought to provide a general mechanism for regulating patterns of gene expression.

Our results indicate that the leucine zipper dimerization interface is present in a functional GCN4 DNA-binding domain; but the α-helix content of GCN4-p (40–43 residues) is greater

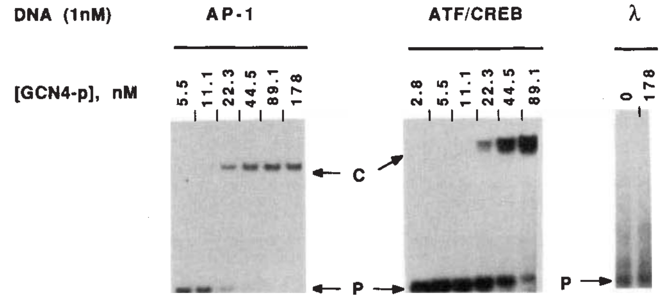

FIG. 1 Gel-retardation assay of GCN4-p binding to oligonucleotides containing the AP-1/TPA-responsive element consensus binding site (5'-GAGATGAGTCATCTC-3'; lanes 1–6), the activating transcription factor/cyclic AMP-responsive element consensus binding site (5'-GAGATGACGTCATCTC; lanes 7–12), or the phage λ O_L1 site (5'-GATACCACTGGCGGTGATATC-3'; lanes 13 and 14). 5'-^{32}P-labelled oligonucleotide (1 nM) was incubated with the indicated concentrations of GCN4-p peptide for 30 min at 23 °C in buffer containing 50 mM KPO_4 (pH 7.5), 50 mM KCl, 100 μg ml^{-1} BSA and 5% glycerol then electrophoresed on a 5% polyacrylamide gel. Following electrophoresis, gels were dried and the free oligonucleotide probe (P) and GCN4-p–DNA complexes (C) were visualized by autoradiography. The GCN4-p peptide (residues 225–281 of the GCN4 protein) was expressed in *E. coli* from a bacteriophage T7 RNA polymerase promoter[24,25] (10–12 mg peptide per litre culture) and purified by a combination of ammonium sulphate precipitation and phosphocellulose chromatography as described elsewhere to >98% homogeneity as judged by SDS–PAGE. An extinction coefficient of 1,313 M^{-1} cm^{-1} at 280 nm, derived by amino acid analysis and ultraviolet spectrophotometry, was used to determine the concentration of GCN4-p monomer in the stock solution. Oligonucleotides used in all assays were synthesized on a Milligen or an Applied Biosystems oligonucleotide synthesizer and purified by HPLC or gel electrophoresis.

than can be accounted for by the leucine-zipper subdomain alone (32–36 residues). Indeed, the CD spectrum of GCN4-p exhibits more α helix as the temperature is lowered. At 1 °C, GCN4-p is ~80% helical, as calculated from a mean residue ellipticity of −27,000 deg cm^2 dmol^{-1} under these conditions (Fig. 2*C*). We propose that the basic region of GCN4-p exists as an ensemble of structures in the absence of DNA, with a substantial population of completely folded molecules present only at low temperature. Support for this proposal comes from studies of a synthetic basic region peptide (GCN4-br; residues 224–249), which exhibits, between 25 °C and 1 °C, a partial α-helical transition similar in magnitude to that of GCN4-p over the same temperature range (Fig. 2*C*). The thermal unfolding transitions of GCN4-p are described in detail elsewhere[13].

The CD spectrum of the specific complex between GCN4-p and a synthetic AP-1 site (21 base pairs) is shown in Fig. 3*A*. A significant increase in the magnitude of the helix-associated bands at 205 and 225 nm is observed for the complex (Fig. 3*A*), relative to GCN4-p alone (Fig. 2*A*). This change is opposite to that expected from the contribution of DNA, which exhibits positive ellipticity in this region of the spectrum (Fig. 2*A*, upper spectrum). The spectral perturbations resulting from complex formation are therefore the result of a DNA-dependent structural transition in the protein, as evidenced by the difference spectrum of the complex minus the sum of its constituents (Fig. 3*A*, Δ). The magnitude of this difference spectrum increases linearly as a function of added AP-1 binding site, reaching saturation at a stoichiometry of two GCN4-p monomers per DNA site (data not shown). At saturation, the DNA-bound conformation of GCN4-p is 95–100% α-helix (Fig. 3*A*). Thus, our results indicate that the folded conformation of the GCN4-p basic region is predominately α-helical, but that the folded state is significantly populated only at low temperature (Fig. 2*C*) or on binding to DNA (Fig. 3*C*). The helical transition observed for the isolated basic region peptide at low temperature (Fig. 2*C*) suggests that these conformational properties of the GCN4

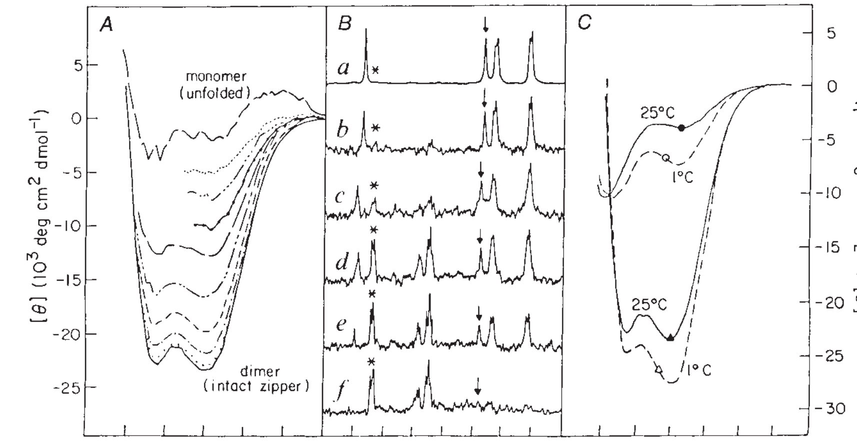

FIG. 2 *A*, normalized far-ultraviolet CD spectra of GCN4-p at different protein concentrations in 200 mM KCl, 50 mM potassium phosphate (pH 7.0 and 25 °C). Protein concentrations (bottom to top, μM): 155, 38, 13, 4.3, 1.6, 1.2, 0.9, 0.6, 0.3, 0.15. Gel-filtration chromatography demonstrates that GCN4-P is dimeric at a protein concentration of 2 mM under these conditions[13]. The GCN4-p monomer seems to be unfolded, as has previously been observed among certain prokaryotic DNA-binding proteins[26]. To maintain adequate signal-to-noise at lower protein concentrations cuvettes with longer path lengths were used (range: 0.5–10 mm). Protein concentration was determined by quantitative amino acid analysis following acid hydrolysis of an aliquot. Mean residue ellipticity [θ] was calculated with calculated molecular weight 6,874 g m^{-1} (118.5 g per mol-residue). *B*, Aromatic region of the 500 Mhz ^{1}H-NMR spectrum of GCN4-p at different protein concentrations in 500 mM KCl, 50 mM potassium phosphate (pD 7.0, direct meter reading) in 99.98% D_2O. Fortuitously, the apparent dimerization constant of GCN4-p is weaker under these conditions, facilitating NMR study. *a–f*, 1,000, 100, 50, 30, 20 and 10 μM, respectively. Dimer-specific and monomer-specific aromatic resonances are observed in slow exchange on the NMR time scale. Arrow indicates dimer-specific Hδ resonance of H266, which decreases in amplitude as the protein concentration is decreased. *, monomer-specific aromatic resonance of Y265, which increases in amplitude as the protein concentration is decreased. The slow-exchange condition provides a lower bound of 10 ms on the lifetime of the dimer. The unresolved resonances of the branched-chain aliphatic sidechains (leucine, isoleucine and valine), which are primarily in the leucine-zipper moiety, are in intermediate exchange under these conditions (data not shown), providing an upper bound of 1 s on the lifetime of the dimer. This short lifetime of the GCN4-p dimer is in apparent conflict with the previous observation suggesting that heterodimers between intact GCN4 and the C-terminal domain were very stable at 0–4 °C in the absence of DNA[5]. But recent experiments suggest that the original observations were artifactual, probably because of co-precipitation of the proteins with excess nonspecific nucleic acid in the particular wheat germ extracts employed at that time. The chemical shift scale was referenced to TSP resonance at 0 p.p.m. *C*, Normalized CD spectra of the GCN4-p dimer (△) and the basic-region peptide (○) at 25 °C and 1 °C in the buffer described in (*A*). The concentration of GCN4-p was 145 μM; the concentration of the basic-region peptide was 154 μM. The sequence of the basic-region peptide is (N)SSDPAALKRARNTEAARRSRARKLQR-$CONH_2$ (single letter amino-acid code). Its structure was verified by amino-acid analysis and sequencing; no truncation or deletion products were observed. The helix content is estimated from the mean residue ellipticity at 222 nm, assuming that for a 100% helical peptide this value is −33,000 deg cm^2 dmol^{-1} at 0 °C, and at higher temperatures is attenuated 0.3% per °C before unfolding[6,27–29] GCN4-p exhibits a cooperative unfolding transition (midpoint 65 °C) similar to that observed in the isolated GCN4 leucine-zipper peptide[27], as will be described elsewhere[13].

basic region are locally determined and are independent of the adjoining leucine-zipper dimerization interface.

Two sets of experiments demonstrate that the folding transition observed on addition of the AP-1 DNA to GCN4-p requires specific complex formation. First, the spectral changes do not occur at high ionic strength (Fig. 3*A*). The CD spectrum of the GCN4-p/AP-1 mixture in 1.5 M KCl is nearly identical to that of GCN4-p alone. The CD spectrum of GCN4-p alone is unaltered in 50–1,500 mM KCl (not shown). Moreover, the KCl-dependence of the DNA-induced spectral changes is consistent with the elution properties of AP-1 site-containing oligonucleotides from a GCN4 affinity column[14]. Second, the CD spectrum of a mixture of GCN4-p with a heterologous DNA control element that does not bind to GCN4-p (the λ O_L1 operator site, Fig. 1) resembles the sum of the spectra of the constituents alone (Fig. 3*B*). The calculated difference spectrum for this mixture (Fig. 3*B*, Δ) has less than 10% of the magnitude of the GCN4-p/AP-1 difference spectrum (Fig. 2*A*). Interestingly, the small perturbation induced by the nonspecific DNA site is consistent with partial stabilization of α-helix.

The DNA-dependent folding of GCN4-p was further examined using a consensus ATF/CREB binding site[15], which differs from the AP-1 site only in the addition of a central CG base pair. In B-form DNA, the two half-sites of this longer element are displaced by a translation of about 3.4 Å and a rotation of 34° relative to their position in the AP-1 site. Although the AP-1 and ATF/CREB sites seem to be targets for distinct families of mammalian leucine-zipper proteins[15], GCN4-p binds to both sites with comparable affinity (Fig. 1). Correspondingly, CD studies of the complex of GCN4-p with an oligonucleotide containing the ATF/CREB site show a folding transition similar to that observed on binding of GCN4-p to the AP-1 sequence (Fig. 3*C*). The CD spectra of both the AP-1 and ATF/CREB DNAs in the region of 250–300 nm are consistent with B-form structure (Fig. 3*D*). Similar, small perturbations in each spectrum are observed on binding GCN4-p. We conclude that the alternative half-site spacings are accommodated by flexibility in GCN4-p rather than by major structural rearrangements in the DNA. Binding to target sites of different length is also observed for *lac* repressor[16], where it is attributed to flexible tethering of individual *lac* headpieces to the core tetramer[17–19].

Our observations by CD spectroscopy of an α-helical folding transition in GCN4-p lead to the following conclusions: (1) the complete GCN4 DNA binding domain undergoes the same dimerization-dependent coil-to-helix transition seen with an isolated leucine-zipper peptide; (2) the basic region of GCN4-p undergoes a coil-to-helix transition when it binds to DNA; (3) α-helical conformation can also be induced in the GCN4-p basic region by lowering the temperature; (4) similar conformational changes are induced in the protein on binding to sites with AP-1 or ATF/CREB consensus elements, implying flexibility in the link between the basic region and the leucine-zipper.

Can our conclusions concerning the coil-to-helix change we have observed when GCN4-p binds DNA be extended to the intact GCN4 protein? We argue strongly that the basic region of intact GCN4 also undergoes such a DNA-dependent folding transition, on the basis of the following observations: (1) GCN4-p binds specifically, with an affinity comparable to that of intact GCN4. Were the basic region of the intact protein 'prefolded,' for example by interaction with segments not present in GCN4-p, its DNA affinity should be higher, corresponding to the free energy expended in folding the peptide; (2) a 60-residue species essentially equivalent to GCN4-p, produced in an *in vitro* translation experiment, binds as a homodimer to a HIS3 target site with the same affinity as a variety of a larger species[4]; (3) in experiments where this 60-residue segment is present with longer chains, large homodimers, heterodimers, and small (60mer) homodimers all bind with the same affinity[5]; (4) our picture of coupled folding and binding is fully consistent with the implications of chemical modification experiments, which show GCN4 and related proteins make extensive contacts with their cognate binding sites[20–22].

These results imply that some part of the DNA-binding domain must be loosely structured to 'wrap' around the DNA helix. One precedent for flexibility in a DNA-binding domain is the N-terminal arm of the λ repressor, which is disordered in the free protein[23]. The arm binds to the 'back side' of the

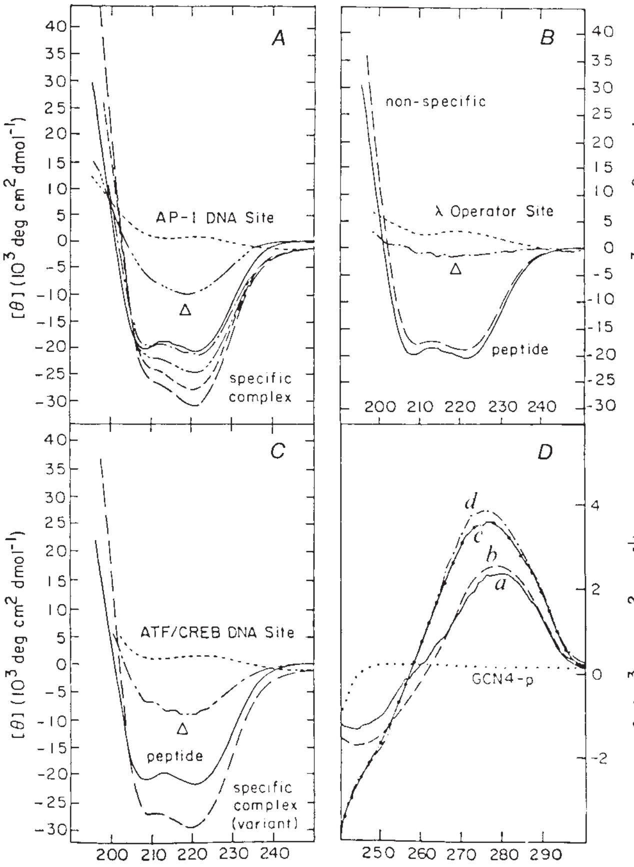

FIG. 3 *A*, Formation of the specific complex with a synthetic AP-1 site results in increased α-helix formation. The GCN4-p concentration was 34 μM. Initial buffer conditions were 50 mM KCl and 12.5 mM potassium phosphate (pH 7.0) at 25 °C. —, the spectrum of GCN4-p is shown; - - -, spectrum of the AP-1 site; — —, spectrum of a 2:1 molar mixture of GCN4-p (one dimer per oligonucleotide duplex) and the AP-1 site (17 μM); —·····—···, the calculated difference spectrum. Δ is calculated as the difference between the observed spectrum of the complex and the sum of the spectra of its constituents. The lineshape and magnitude of Δ is in accord with a coil-to-helix transition[30]. Dissociation of the specific complex occurs at higher concentrations of KCl: spectra are shown at 500 mM KCl (— —), 700 mM KCl (—-—) and 1,500 mM KCl (—-—). The 21-base-pair AP-1 site has sequence GAGATGAGTCATCTC; a similar DNA-dependent transition is observed with a consensus 15-base-pair AP-1 site. *B*, An unrelated control DNA site (the λ operator site O_L1) does not appreciably affect the helix content of GCN4-p. —, the spectrum of GCN4-p; - - -, the spectrum of the λ operator; — —, the spectrum of a 2:1 molar mixture of GCN4-p and the λ site; —·—·, the calculated difference spectrum (defined in *A*). The sequence of the λ operator site is GATACCACTGGCGGTGATATC and its complement. *C*, The ATF/CREB DNA site, which contains an additional GC base pair in the centre of the recognition site, induces a similar increase in helix content on formation of a specific complex. —, the spectrum of GCN4-p; - - -, the spectrum of the ATF/CREB site (GAGATGA$\underline{C}$GTCATCTC) additional central base-pair underlined; the — —, the spectrum of a 2:1 molar mixture of GCN4-p (one dimer per oligonucleotide duplex) and the ATF/CREB site —·····—···, the calculated difference spectrum (Δ). *D*, Conformation of the AP-1 (spectrum *a*; 15-base-pair sequence described in *A*) and ATF/CREB (spectrum *b*; related 16-base-pair sequence shown in *C*) oligonucleotides may be observed in the near ultraviolet region of the CD spectrum. The ellipticity of GCN4-p in this region (dotted line) is near zero. The spectrum of the AP-1 complex (*c*) and ATF/CREB complex (*d*) show similar small changes in amplitude and position of the local maximum.

DNA helix and forms contacts that, in part, determine the operator specificity of the repressor[23]. In GCN4, the entire recognition region seems to undergo a folding transition on binding DNA. This can be a rather general mechanism for reducing the kinetic barriers in the assembly of protein-nucleic acid complexes.

Our results do not specify the details of α-helical packing in the basic region of GCN4. In the scissors grip model proposed for this class of DNA-binding proteins[20], the basic region forms an α-helical extension of the leucine-zipper, with a kink near the N-terminus. Although consistent with such a model structure, our data are also consistent with an array of α-helical segments within the basic region. Our data do, however, show that the basic region of GCN4 is almost completely α-helical when bound to DNA and that any kinks or turns in the peptide backbone involve very few residues. □

Received 30 May; accepted 15 August 1990.

1. Pabo, C. O. & Sauer, R. T. *A. Rev. Biochem.* **53,** 293-321 (1984).
2. Klug, A. & Rhodes, D. *Trends Biochem. Sci.* **12,** 464-467 (1987).
3. Landschulz, W. H., Johnson, P. F. & McKnight, S. L. *Science* **243,** 1681-1686 (1989).
4. Hope, I. A. & Struhl, K. *Cell* **43,** 177-188 (1985).
5. Hope, I. A. & Struhl, K. *EMBO J.* **6,** 2781-2784 (1987).
6. O'Shea, E. K., Rutkowski, R. & Kim. P. S. *Science* **243,** 538-542 (1989).
7. Hope, I. A. & Struhl, K. *Cell* **46,** 885-894 (1986).
8. Oas, T. G., McIntosh, L. P., O'Shea, E. K., Dahlquist, F. W. & Kim, P. S. *Biochemistry* **29,** 2891-2894 (1990).
9. Kouzarides, T. & Ziff, E. *Nature* **340,** 568-571 (1989).
10. Sellars, J. W. & Struhl, K. *Nature* **341,** 74-76 (1989).
11. Landschultz, W. H., Johnson, P. F. & McKnight, S. L. *Science* **246,** 922-926 (1988).
12. Agre, P., Johnson, P. F. & McKnight, S. L. *Scinece* **246,** 922-926 (1989).
13. Weiss, M. A. *Biochemistry* (in the press).
14. Oliphant, A. R., Brandl, C. J. & Struhl, K. *Molec. cell. Biol.* **9,** 2944-2949 (1989).
15. Hai, T., Liu, F., Allegretto, E. A., Karin, M., & Green, M. R. *Genes Dev.* **2,** 1216-1226 (1988).
16. Sadler, J. R., Sasmor, H. & Betz, J. L. *Proc. natn. Acad. Sci. U.S.A.* **80,** 6785-6789 (1983).
17. Arndt, K., Boschelli, F., Lu, P & Miller, J. H. *Biochemistry* **20,** 6109-6118 (1981).
18. Buck, F., Ruterjans, H. & Beyreuther, K. *FEBS Lett.* **96,** 335-338 (1978).
19. Wade-Jardetzky, N. *et al. J. molec. Biol.* **128,** 259-264 (1979).
20. Vinson, C. R., Sigler, P. B. & McKnight, S. L. *Science* **246,** 911-916 (1989).
21. Gartenberg, M. R., Ampe, C., Steitz, T. A. & Crothers, D. M. *Proc. natn. Acad. Sci. U.S.A.* **87,** 6034-6038 (1990).
22. Oakly, M. G. & Dervan, P. B. *Science* **248,** 847-850 (1990).
23. Jordan, S. & Pabo, C. O. *Science* **242,** 895-899 (1988).
24. Rosenberg, A. H. *et al. Gene* **56,** 125-135 (1987).
25. Studier, F. W. & Moffat, B. A. *J. molec. Biol.* **189,** 113-130 (1986).
26. Bowie, J. U. & Sauer, R. T. *Biochemistry* **28,** 7139-7143 (1989).
27. O'Shea, E. K., Rutkowski, R., Stafford, W. F. III & Kim, P. S. *Science* **243,** 1689-1694 (1989).
28. Lehrer, S. S., Qian, Y. & Hvidt, S. *Science* **246,** 926-928 (1989).
29. Chen, Y.-H., Yang, J. T. & Chou, K. H. *Biochemistry* **13,** 3350-3356 (1974).
30. Johnson, W. C. Jr *Protein Secondary Structure and Circular Dichroism: a Practical Guide* 205-214 (1990).

ACKNOWLEDGEMENTS. We thank S. Leher (Boston Biomedical Research Institute) for discussion and use of his CD spectropolarimeter; H.T. Keutmann and S. Magil (Massachusetts General Hospital) for peptide characterization; K.A. Mason for technical assistance; and J. Habener, H. Kronenberg, and members of our laboratories for helpful discussion. This work was supported by the NIH (M.A.W., S.C.H. and K.S.) and the Markey Charitable Trust at Harvard Medical School and by the Howard Hughes Medical Institute. M.A.W. is supported in part by the Pfizer Scholars Program for New Faculty and the American Cancer Society. T.E.E. is supported by a postdoctoral fellowship from the NIH. C.R.W. is supported by a postdoctoral fellowship from Merck, Sharp & Dohme. The NMR spectrometer at Harvard Medical School is supported by NIH shared Instrumentation grant 516RR04862-0 and the National Health Resources Foundation.

Phylogenetic and genetic evidence for base-triples in the catalytic domain of group I introns

François Michel*, Andrew D. Ellington, Sandra Couture & Jack W. Szostak†

Department of Molecular Biology, Massachusetts General Hospital,
Boston, Massachusetts, 02114, USA

UNDERSTANDING the mechanisms by which ribozymes catalyse chemical reactions requires a detailed knowledge of their structure. The secondary structure of the group I introns has been confirmed by comparison of over 70 published sequences[1–4], by chemical protection studies[5], and by genetic experiments involving compensatory mutations[2,6,7]. Phylogenetic data can also be used to identify tertiary interactions in RNA molecules. This was first done by Levitt[8], who predicted tertiary interactions in transfer RNA, which were subsequently confirmed by X-ray crystallography[9]. More recently, sequence comparison data have been used to predict tertiary interactions in ribosomal RNA[10]. We have searched a complete alignment of the core regions of group I introns[1,2] for evolutionary covariations that could not be ascribed to classical Watson–Crick or wobble base pairings. Here we describe two examples of phylogenetic covariation that are most simply explained by postulating hydrogen-bonded base-triples similar to those found in tRNA. Genetic experiments with the *Tetrahymena* and *sunY* introns confirm the importance of these interactions for the structure of the ribozyme.

Comparison of group I intron sequences reveals a striking covariation of the second and third base pairs of stem P4 and the first and second nucleotides, respectively, of the single-stranded segment J6/7 (see Fig. 1*a*). Base pair two of stem P4 is most often G·C or C·G; when P4-2 is G·C, the first base in J6/7 is always U; when P4-2 is C·G, J6/7-1 is usually G, less often C, but never U. Several independent examples of concerted changes at these positions are seen in phylogeny (Fig. 1*b*), indicating a functional requirement as opposed to mere evolutionary drift. For example, G·C-U changes to C·G-G within both distinct subclasses of the group I introns, showing that this triple mutation has occurred at least twice during their divergence. Similarly, P4 base pair 3 is usually C·G or G·C, and these base pairs are associated with C or U, respectively, at position J6/7-2; again, independent concerted changes of these positions are seen in phylogeny.

Comparing closely related introns can aid in distinguishing between sequence changes that are due to functional necessity as opposed to evolutionary distance. Two bacteriophage T4 introns, T4 *td* and T4 *sunY*, are virtually identical within the conserved core of group I introns[11], yet differ at the second and third base pairs of P4 and the second nucleotide of J6/7. It is unlikely that these sequences would have covaried if the base substitutions were merely the product of neutral drift.

To assess the chemical feasibility of the interactions detected by covariation, we have examined physical models of the proposed triples. All of the phylogenetically related base triples can be modelled as geometrically equivalent, or near-equivalent, structures (Fig. 2). In most cases, hydrogen bond donors and acceptors are interchanged without affecting the geometry of the interaction. The base-triples we propose between P4 and J6/7 are similar to those seen in tRNA[8,9], in that a single base interacts with one base of a Watson–Crick base pair on the major groove face of the base pair, with all three bases being roughly coplanar.

The small number of base changes between the bacteriophage T4 introns *td* and *sunY* made these introns ideal for initial tests of our model. If P4 and J6/7 do interact in a functionally significant manner, changes in one or the other of these regions should be deleterious, whereas changing both together should restore the original structure and enzymatic activity. We started with *sunY* and constructed all possible intermediates on the way to a *td*-like ribozyme, varying J6/7 and the base pairs of P4. The splicing activity of these intermediate forms was measured in the presence of saturating guanosine and the initial velocity of the reaction was determined (Fig. 3). Both intermediates that replace one base triple from *sunY* with the *td*-like base triple are slightly more active than the *sunY* wild type. In contrast, all mutants in which a predicted base-triple is disrupted have decreased activity. Loss of a hydrogen bond from a predicted base-triple (for example, C·G-C to C·G-U) reduces activity

* Permanent address: Centre de Génétique Moléculaire du CNRS, 91190 Gif-sur-Yvette, France.
† To whom correspondence should be addressed.