1995

DNA Recognition by Helix-Loop-Helix Proteins

Adrian Ferre-D'Amare

DNA Recognition by Helix-Loop-Helix Proteins

Adrian R. Ferré-D'Amaré

Thesis submitted to the Faculty of The Rockefeller University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Preface

I came to The Rockefeller University in August of 1990 with an interest in molecular recognition in biochemistry, in how molecules self-assemble into the supra-molecular complexes that result in life. This Dissertation is a summary of four years of work elucidating the structural underpinnings of a biologically important recognition process: sequence-specific DNA binding by the Helix-Loop-Helix proteins, a family of eukaryotic transcription factors.

None of the work described here would have been accomplished without innumerable contributions from the scientists I have been associated with. I wish to thank my advisor, Dr. Stephen Burley, for his unwavering support. I am grateful to Drs. Kirk Clark, Joseph Kim, and Xiang-Peng Kong for valuable discussions and for their help with many experiments. I wish to thank Dr. Magda Konarska for accepting me as a rotation student, back in 1990. The five months I spent in her laboratory I remember as a period of

extraordinarily intense and enjoyable learning. I am grateful to my collaborators, Drs. Philippe Pognonec and Robert Roeder on Upstream Stimulatory Factor, Drs. George Prendergast and Edward Ziff on Max, Mr. Steven Cohen and Dr. Brian Chait on the application of mass spectrometry to the elucidation of biomolecular structure and function, and Drs. Lynne Canne and Stephen Kent on the synthesis of covalently linked b/HLH/Z dimers. I appreciate the advice given to me on CD spectroscopy by Dr. David Cowburn, and on some aspects of X-ray crystallography by Dr. John Kuriyan. I thank Dr. David Mauzerall, chair of my thesis committee, and Dr. Paul Sigler, my external examiner, for their valuable time.

I am grateful to The Rockefeller University for full financial support, in the form of a Rockefeller University Graduate Fellowship and a David Rockefeller Predoctoral Fellowship. I wish to thank the *de facto* Deans of Graduate Studies, Dr. Mary Rifkin, and after her untimely departure, Dr. Peter Model, for their valiant efforts to safeguard the unique Graduate Program of The Rockefeller University.

Looking back on my years as a graduate student, I am happy to be able to say that the means justified the end. The company of my colleagues, friends, relatives, and parents made the voyage more important than the destination. To them all, I am very grateful.

New York City, November 1994

Contents

List of Figures viii

List of Tables xi

Abbreviations xiii

Abstract 1

Chapter 1 Introduction:

Sequence-Specific DNA-Binding Proteins 2

1.1 The Regulation of Gene Expression by DNA-Binding Proteins 3

1.2 Structural Studies on DNA-Protein Interactions 15

1.3 Helix-Loop-Helix Transcription Factors 26

1.5 Aims of this Work 34

Chapter 2 Materials and Methods 36
2.1 Protein Purification 37
2.2 DNA Purification 48
2.3 Electrophoretic Mobility Shift Assays 49
2.4 Circular Dichroism Spectroscopy 51
2.5 Dynamic Light Scattering 52
2.6 Analytical Ultracentrifugation 54
2.7 Bimolecular Ligation 55
2.8 Crystallization 56
2.9 X-ray Diffraction Data Collection 62
2.10 Structure Determination and Crystallographic Refinement 64

Chapter 3 Results 65
3.1 Biochemical Characterization of DNA-Binding by USF 66
3.2 Crystal Structure Determination of the Max b/HLH/Z Domain Dimer in Complex with DNA 82
3.3 Structure of the Max b/HLH/Z Domain Dimier in Complex with DNA 106
3.4 Structure of a USF b/HLH Dimer Bound to DNA 124
3.5 Variations on the HLH Theme: Work in Progress 133

Chapter 4 Discussion 145
4.1 A New Protein Fold: Dimerization of HLH Proteins 146
4.2 DNA Recognition by HLH Proteins 162

4.3 Biological Function of HLH Proteins 167

4.4 Future Directions 177

Appendix Use of Dynamic Light Scattering to Assess Crystallizability of Macromolecules and Macromolecular Assemblies 179

References 184

List of Figures

Chapter 1 Introduction: Sequence-Specific DNA-Binding Proteins

Figure 1. Representative examples of sequence-specific DNA binding proteins whose three-dimensional structures in complex with DNA have been determined 21

Figure 2. Partial sequence alignment of representative basic/helix-loop-helix/leucine zipper (b/HLH/Z), basic/helix-loop-helix (b/HLH), helix-loop-helix proteins lacking a basic region (HLH) 29

Chapter 2 Materials and Methods

Figure 3. Amino acid sequences in one letter code of proteins prepared for this study 39

Figure 4. Sequences of double-stranded DNA oligonucleotides which yielded diffraction quality HLH cocrystals 59

Chapter 3 Results

Figure 5. Determination of protein-DNA stoichiometries and DNA-binding activity and specificity for USF and USF b/HLH/Z 67

Figure 6. Normalized CD spectra of USF constructs 71

Figure 7. Hydrodynamic characterization of some USF constructs 75

Figure 8. Bimolecular ligation experiment 80

Figure 9. EMSA and CD characterization of Max constructs 83

Figure 10. Hydrodynamic characterization of Max constructs 86

Figure 11. Sample HLH-DNA cocrystals and diffraction patterns 89

Figure 12. The w = 1/6 Harker section of the isomorphous difference Patterson synthesis for derivative IdU4R 94

Figure 13. Final electron density and crystal packing of the (Max22-113)$_2$-DNA complex structure 100

Figure 14. Ramachandran plot of the final refined (Max22-113)$_2$-DNA complex 102

Figure 15. Luzzati plot of the final refined (Max22-113)$_2$-DNA complex 103

Figure 16. Individual isotropic B-factors of the final refined (Max22-113)$_2$-DNA complex..... 104

Figure 17. Overall view of the structure of the (Max22-113)$_2$-DNA complex 107

Figure 18. Solvent accessible surface areas buried upon interaction of various parts of the refined (Max22-113)$_2$-DNA complex 110

Figure 19. Some stereochemical parameters of the DNA in complex with Max 114

Figure 20. Parallel-eye stereo representations of portions of the (Max22-113)$_2$-DNA complex..... 118

Figure 21. Max basic region-E-box interactions 122

Figure 22. Some salient features of the USF (b/HLH)$_2$-DNA structure 131

Figure 23. Normalized CD spectra of USF b/HLH, c/sHRY and c/sSREBP 135

Figure 24. Biochemical characterization of synthetic b/HLH/Z dimers 141

Chapter 4 Discussion

Figure 25. b/HLH/Z dimerization and DNA recognition 148

Figure 26. Three-dimensional alignment of the Max b/HLH domain with E47 and MyoD 152

Figure 27. Comparison of the Max b/HLH/Z-DNA complex with the GCN4 b/Z-DNA complex 159

List of Tables

Chapter 1 Introduction: Sequence-Specific DNA-Binding Proteins

Table 1. Structures of DNA-protein complexes published before November 1994 18

Chapter 2 Materials and Methods

Table 2. Typical parameters varied for firtst factorial screen of new DNA-HLH complexes 57

Chapter 3 Results

Table 3. Summary of DLS results from monodisperse samples 78

Table 4. Data statistics 91

Table 5. Phasing statistics 95

Table 6. Refined heavy atom parameters 97

Table 7. Crystallographic refinement statistics of the (Max22-113)$_2$-DNA complex 97

Table 8. Statistics of merged USF (b/HLH)$_2$-DNA diffraction data 126

Table 9. Crystallographic refinement statistics of the (b/HLH)$_2$-DNA complex 126

Abbreviations

The one-letter code for the twenty amino acids is given in the legend to Fig. 2

A	Activation domain (in the context A/b/HLH)
a.m.u.	Atomic mass unit
b	Basic region (in the context b/HLH or b/HLH/Z)
Bis-Tris Propane	1,3-Bis[tris-(hydroxymethyl)-methylamino]propane
DTT	Dithiothreitol (*threo*-1,4-dimercapto-2,3-butanediol)
EDTA	Ethylenediaminetetraacetic acid
IPTG	Isopropyl-β-D-thiogalactopyranoside
Hepes	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HLH	Helix-Loop-Helix
MALDI	Matrix-assisted laser desorption ionization

Mes	4-Morpholineethanesulfonic acid
MIR	Multiple isomorphous replacement
MPD	2-Methyl-2,4-pentanediol (hexylene glycol)
NP-40	Nonidet P-40 (1,1,3,3-tetramethylbutylphenyl polyethyleneoxide; non-ionic detergent)
octyl glucoside	1-O-Octyl-β-D-glucopyranoside
PDB	Brookhaven Protein Data Bank
PEG	Polyethyleneglycol
PEG-MME	Polyethyleneglycol monomethylether
PEI	Polyethyleneimine
PMSF	Phenylmethylsulfonylfluoride (α-Toluenesulfonyl fluoride)
r.m.s.d.	Root-mean-square deviation
Tris	Tris(hydroxymethyl)aminomethane
Z	Leucine zipper (in the context b/Z or b/HLH/Z)

Abstract

The Helix-Loop-Helix (HLH) family of eukaryotic transcription factors comprises a large number of proteins which play key roles in homeostasis, the regulation cell proliferation and differentiation. These proteins share a phylogenetically conserved bipartite b/HLH domain responsible for specific DNA binding and dimerization. The HLH region dictates dimerization affinity and specificity while the basic region (b) is primarily responsible for sequence-specific DNA binding. In some family members, such as the Myc oncoproteins, the HLH motif is followed by a heptad repeat of hydrophobic amino acids, or "leucine zipper" (Z). My X-ray crystallographic structure determination at 2.9Å resolution of a dimer of the b/HLH/Z domain of the mammalian oncoprotein Max bound to its target DNA revealed that this symmetric homodimer folds into a novel parallel left-handed four-helix bundle, which is globular and stabilized by a well-defined hydrophobic core. Two pairs of α-helices protrude in opposite directions from the bundle. One, the basic regions, enters the major groove of the target B-form DNA, and makes numerous contacts with the bases and phosphodiester backbone. The other, the leucine zipper, forms a left-handed coiled coil, extending the hydrophobic interface of the homodimer. I also determined the cocrystal structure of a truncated b/HLH homodimer of the human transcription factor USF bound to DNA. As expected from the sequence conservation, this protein adopts the same three-dimensional structure as Max b/HLH. Circular dichroism spectroscopic investigation of DNA binding by Max and USF demonstrated, in concert with their cocrystal structures, that these proteins undergo a dramatic folding transition upon specific, high-affinity DNA binding. More than forty residues per dimer become α-helical upon association. I also demonstrated by hydrodynamic as well as biochemical methods that these proteins can form bivalent tetramers at physiologically meaningful concentrations. This suggests that they may play a role in DNA looping, thought to be important in the transcriptional regulation of eukaryotic genes.

Chapter 1

Introduction: Sequence-Specific DNA-Binding Proteins

Section 1.1 reviews the historical development of ideas on the role that sequence-specific DNA-binding proteins play in the regulated expression of genetic information. This starts with a review of the notion of the *repressor*. A summary of investigation on the affinity, specificity, and kinetics of its association with DNA follows. Then some key features of the transcriptional regulation of gene expression in eukaryotes are reviewed. Starting in the mid-1980's, a significant quantity of structural information has become available on the way in which proteins belonging to different families recognize specific DNA sequences. This information is reviewed in Section 1.2. In order to keep the length of the Introduction reasonable, DNA modifying enzymes are not considered. Section 1.3 reviews biochemical and biological background on the proteins that constitute the subject of this dissertation, the Helix-Loop-Helix proteins. Background material additional to that covered in Sections 1.2

and 1.3 is reviewed in context in the Discussion (Chapter 4). Finally, Section 1.4 defines the aims of this work.

1.1 Regulation of Gene Expression by DNA-Binding Proteins

The Repressor The operon hypothesis of Jacob and Monod (Jacob and Monod, 1961) postulated the existence of a diffusible molecular entity, the *repressor*, which would regulate a group of genes by recognizing a control element present in *cis* to them, the *operator*. The repressor was put forward in order to account for "specific pleiotropy" of a regulatory gene, that is, for its ability to repress the activity of a defined, small set of non-cistronic loci. For instance, the product of the *i* gene (the Lac Repressor) had been shown to repress co-ordinately the expression of three activities: β-galactosidase, galactoside-transacetylase, and a then hypothetical galactoside-permease. Importantly, this repressive activity displayed a stereospecificity for inducer different from that shown for their enzymatic substrates by either of the then characterized activities which it controlled. Mutations in the β-galactosidase gene failed to affect the galactoside-transacetylase and conversely; mutations in the *i* gene, however, affected the regulation of *both* activities (thus being *pleiotropic*) without appearing to affect any other gene in the bacterium (thus being *specific*). The stereospecificity for inducer suggested the participation of some kind of protein in the repressive activity (although certain experiments mislead Jacob and Monod to propose that the repressor was more likely to be RNA) and its locus of action was proposed to be either the gene (DNA) or the transcript (itself a novel concept).

The biochemical breakthroughs that opened the way to understanding the chemistry underlying the mode of action of the repressor came in 1966 when the Lac Repressor (Gilbert and Müller-Hill, 1966) and the λ Repressor (Ptashne, 1967a) were partially

purified and shown to be *proteins* of molecular weights of *ca.* 200 and 30 kDa, respectively. The Lac Repressor was purified exploiting its affinity for a radioactive inducer (IPTG); λ by differential double isotopic labeling of cells infected with phage capable or incapable of producing repressor followed by biochemical fractionation of cell extracts and assay of chromatographic fractions for enrichment in one but not the other isotope. The biochemical availability of the repressors allowed these investigators to show that these proteins bound tightly to phage DNA containing wild-type operators but only weakly or undetectably to phage DNA with mutant operators. Lac Repressor dissociated from operator DNA when IPTG was added to the reaction. Furthermore, the binding of both repressors was obliterated when the operator DNA was denatured (Ptashne, 1967b; Gilbert and Müller-Hill, 1967). It had been demonstrated that *"The ... operator is DNA"*, and that repressor proteins bound directly and specifically to it. It was subsequently demonstrated that these proteins repress transcription when bound to their operators because the operators overlap with the promoter, the DNA element that recruits RNA polymerase. Repression occurs through direct physical competition for binding (Squires *et al.*, 1975). These discoveries made the sequence-specific binding of proteins to DNA a central subject of study in molecular biology.

The genetic information stored in DNA is transcribed by RNA polymerase into (messenger) RNA, which is then used during translation as a template for protein synthesis. The regulation of the expression of genetic information could, therefore, occur either during transcription or translation, by modulating the steady-state levels of the products of transcription or translation (RNA or protein, respectively), or by modulating the biochemical activity of these products. Not unexpectedly, all possible levels of regulation are employed by living organisms. Nonetheless, regulation at the level of initiation of transcription plays a preponderant role in the majority of genes which have been examined (reviewed, *e.g.* in Mitchell and Tjian, 1989)

Affinity and Specificity Ptashne (1967b) estimated the dissociation constant of λ Repressor for operator DNA to be of the order of 10^{-10} M; Gilbert (1967) estimated that of the Lac Repressor for its operator to be of the order of 10^{-12} M. Ptashne and Gilbert also noticed that the affinities of the repressors for DNA decreased considerably when the ionic strength of the solutions was increased, implying that electrostatic interactions accounted for part of the binding free energy (see, *e.g.*, Misra *et al.*, 1994). Curiously, the affinity of the Lac Repressor for operator DNA implied that, if its rate of association with operator was diffusion limited, then its off-rate would be several hours, much longer that the time needed for induction to occur *in vivo*.

These findings were soon followed by more detailed characterizations of the chemistry of the repressor-operator interactions. In a series of pioneering contributions, Riggs, Bourgeois, and coworkers (Riggs and Bourgeois, 1968; Riggs *et al.*, 1968; Riggs *et al.*, 1970a; Riggs *et al.*, 1970b; Riggs *et al.*, 1970c; reviewed in Barkley and Bourgeois, 1978) developed and exploited the filter binding assay to address a number of fundamental questions including (1) affinity of the Lac Repressor for operator DNA, (2) affinity of repressor for non-operator DNA, (3) specificity, defined as the ratio of affinities for specific (operator) and non-specific (non-operator) DNAs, (4) kinetics of DNA binding, (5) enthalpic and entropic contributions to the free energy of binding (through van't Hoff analysis), and (6) effects of pH and ionic strength on binding. *Inter alia*, they found that the speed with which the repressor found the operator appeared to exceed the limits set by diffusion, and that binding at 24°C was entropically driven: their results are ΔG = -18 kcal mol^{-1}, ΔH = + 8.5 kcal mol^{-1}, ΔS = + 90 cal $mol^{-1}K^{-1}$.

Repressors and other sequence-specific DNA-binding proteins have a finite affinity for non-specific DNA, that is, DNA that has a sequence different from their target(s).

Specificity can operationally be defined as the ratio of specific to non-specific equilibrium association constants. (The information content of a specific binding site is discussed below.) *In vivo* measurements in *E. coli* for the Lac Repressor gave values for the dissociation constant from operator of 10^{-12} M and from non-specific DNA of 10^{-9} M (Kao-Huang *et al.*, 1977). This 1000-fold specificity of binding dictates the necessity of having more repressor molecules in the bacterium than would be required with "infinite" specificity in order to achieve any given occupancy of the repressor. These measurements also implied that the ionic strength (in NaCl equivalents) of the cytosol of the bacterium was between 0.17 and 0.24 M, potentially establishing the relevant ionic strength for performing affinity measurements *in vitro*. However, It was shown later that the chemical nature of the anion had a profound effect on the association of DNA-binding proteins to their targets (Leirmo *et al.*, 1987) with *in vitro* association rate constants varying by up to 30-fold depending on whether potassium glutamate or chloride was employed. The non-specific binding of sequence-specific DNA binding proteins has important implications for the kinetics of association with the specific targets, as will be mentioned below.

DNA is a highly hydrated molecule, with approximately twenty water molecules associated with each nucleotide under physiologic conditions (reviewed in Saenger, 1984). Furthermore, its high negative electric charge density results in a phenomenon known as "counter-ion condensation": regardless of the bulk concentration of cations, the concentration of these positive counter-ions in the immediate neighborhood of DNA is rather constant, with approximately 80% of the charge of the phosphate backbone neutralized (reviewed in Manning, 1978). The patterns of both hydration and counter-ion binding (*e.g.* Buckin *et al.*, 1994) depend to some extent on the sequence of the DNA. Riggs and Bourgeois (1970) suggested that the entropic driving force for Lac Repressor-operator association could be a consequence of water and cation displacement, and further, that " ... it is quite possible that specificity results only from a lack of steric hindrance,

given the correct base sequence." Some recent genetic results can be taken to support such a view (Lehming *et al.*, 1990). Microcalorimetry (Takeda *et al.*, 1992) and DNaseI footprint titrations (Senear and Batey, 1991) implied that cation and water release is responsible for both operator and non-operator DNA binding by repressor; another study (Koblan and Ackers, 1991a; Koblan and Ackers, 1991b), using the same footprint methodology, arrived at the conclusion that the ability of the repressor (λ in this case) to discriminate between various high and low affinity operators is linked to differential ion binding/release. In the absence of detailed structural knowledge of the free and bound components one cannot exclude conformational changes in either or both repressor and operator to account for the net entropic increase (Riggs *et al.*, 1970).

Takeda *et al.* (1992) found by titration microcalorimetry that the association of Cro protein from bacteriophage λ to one of its high affinity operators proceeded with ΔG = -16.1 kcal mol^{-1}, ΔH = + 0.8 kcal mol^{-1}, ΔS = +59 cal mol^{-1}K^{-1}, and ΔC_p = -360 cal mol^{-1}K^{-1} (at 15°C and 0.1 M supporting electrolyte). For non-specific association, the thermodynamic parameters were ΔG = -9.7 kcal mol^{-1}, ΔH = + 4.4 kcal mol^{-1}, ΔS = +49 cal mol^{-1}K^{-1}, and $\Delta C_p \approx 0$ (1 cal = 4.18 J). Both kinds of associations were thus entropy driven. The absence of change in heat capacity and the large positive entropy change during non-specific binding imply a lack in perturbation of enthalpic states of the molecules participating in complex formation, and release of a counterions and bound water; this association was characterized as "loose" and mostly electrostatic (driven entropically by ion and water release). The large negative change in heat capacity associated with specific binding was interpreted to imply a "narrowing" of enthalpic states of the complex (or "tightening" of the structure) as a result of the formation of further protein-DNA interactions (hydrogen-bonds, van der Waals contacts, electrostatic interactions) and concomitant release of more ions and water.

It has been established for several decades now that the large negative heat capacity change observed upon protein folding is a signature of burial of non-polar surface area in reactions taking place in aqueous medium (Sturtevant, 1977; Kuntz and Kauzmann, 1974; Baldwin, 1986; Spolar *et al.*, 1989). Burial of non-polar surface area, the hydrophobic effect, is thought to be the main driving force behind protein folding. The loss of conformational entropy resulting from folding an extended polypeptide chain into a compact structure that can explore many fewer conformational states is the main opposing force. In this view, the extensive intramolecular hydrogen bonding and other polar interactions seen in the interior of protein structures result in little favorable enthalpy of folding, since for each hydrogen bond formed within the folded protein, two hydrogen bonds between the unfolded protein and water must be disrupted. The hydrogen bonds present within protein structures arise from a need to maintain buried partial charges neutralized in the hydrophobic interior of a protein that resulted from "hydrophobic collapse" (reviewed in Dill, 1990). The large negative heat capacity change observed upon protein folding can also result from loss in the number of "soft" vibrational modes accessible to the protein upon folding (Sturtevant, 1977). Observation of a large negative heat capacity change upon protein-DNA association led Ha *et al.* (1989) to propose, by analogy, that burial of non-polar surface area as well as release of cations ware the main driving forces behind protein-DNA association.

Investigation of the thermodynamics of Trp Repressor-operator interaction by titration microcalorimetry (Ladbury *et al.*, 1994) showed that unlike the other repressor-operator associations examined calorimetrically, this one is enthalpy driven throughout the physiologic temperature range. Two modes of Trp Repressor binding to the operator could be detected, and the second one was associated with a large negative heat capacity change, but the abundant structural information available on this system precludes any explanation in terms of large folding transitions concomitant with binding. The authors attributed the

heat capacity change to "tightening" of the structure, to reduction in vibrational modes accessible to the complex.

If protein folding and oligomerization are driven by the same forces responsible for specific DNA-protein association, a survey of the chemical nature of protein-protein interfaces might be useful. Inspection of a number of homo- or hetero-oligomeric structures led to the conclusion that, first, the same relationship between buried non-polar solvent accessible area and folding free energy per unit molecular weight of protein (a way of quantitating the hydrophobic "force") found to operate within protein molecules (Eisenberg and McLachlan, 1986) applied to oligomer interfaces (Miller *et al.*, 1987). Second, the chemical nature of protein-protein interfaces was seen to be highly variable, with hydrophobic interactions predominating in some complexes, while polar (hydrogen bonding and ion-pair formation) interactions played a more significant role in others (Janin and Chothia, 1990). The average interaction surface of 19 proteins examined by these authors did not differ from the solvent accessible surface of an average protein: 55% non-polar, 25% polar, and 20% charged. Overall, no particularity was found in the amino acid composition of the interfaces.

Information Content of Operators The availability of the nucleotide sequences of operators (Gilbert and Maxam, 1973) made it possible to define specificity of repressor-operator interactions as a function of DNA sequence. Schneider *et al.* (1986) proposed two measures of the information content of binding sites on nucleotide sequences. R_{sequence} is a measure of the information in the sequence patterns of binding sites. It is derived from an alignment of different sites bound by the protein under study employing information theoretical arguments derived from statistical mechanics, *i.e.*, by treating uncertainty as entropy. The uncertainty H for a position where there are M possible residues (4 for DNA) each occurring with a probability P is:

$$H = -\sum_{i=1}^{M} P_i \log_2 P_i$$

For a random DNA sequences each P is 1/4 and thus the uncertainty H_r is 2 bits per position in the sequence. For non-random sequences, the probability $P(B,L)$ of finding a base B in each position L is worked out from a sequence alignment, and the uncertainty of the sequence H_s at position L is then:

$$H_s(L) = -\sum_{B=A,T,C,G} P(B,L)\log_2 P(B,L)$$

The information content at a site is:

$$R_s(L) = H_r - H_s(L)$$

For a sequence of length n, the information content in bits is,

$$R_s = \sum_{i=1}^{n} R_s(i) = \sum_{i=1}^{n} (H_r - H_s(L))$$

For multi-partite binding sites, the uncertainty in the spacing can also be calculated from sequence alignments, and this uncertainty can be subtracted from the sum of the information content of the two sites.

The same authors also derived a measure of the amount of information that would be required to find a site in a given genome, $R_{\text{frequency}}$, which is the difference (in base 2 logarithms): total number of bases in the genome minus total number of sites in the genome. Interestingly, in the case of the bacterial operators examined by those authors, the two measures gave similar information content, suggesting that binding sites evolve to have just enough information to be located. For instance, R_{sequence} was 19.3, 20.6, and 17.5 bits per site for Lac, Trp, and λ Repressor or Cro, respectively while $R_{\text{frequency}}$ was calculated to be 20.9, 20.3, and 19.3 bits per genome for the same proteins. These measures of information content are useful in pinpointing which positions in an alignment

of binding sites contain the most information overcoming the limitations of selecting a "consensus" sequence out of a population of recognized sites, and in analyzing whether a given site contains enough information relative to what is required to locate it, but do not restrict the physical-chemical mechanism of recognition. Also, the analysis supposes that no relevant information is contained outside of the locus being considered.

Kinetics of DNA-Repressor Association The association rate of repressor to an operator embedded in genomic DNA was observed to exceed the diffusion limit. This was suggested by several groups to result from "facilitated diffusion" (Adam and Delbrück, 1968; Richter and Eigen, 1974; Berg *et al.*, 1981; Winter and von Hippel, 1981; Winter *et al.*, 1981). Since repressor has a finite affinity for non-operator DNA within a huge excess of which the operator is located, it was proposed that the repressor would bind to any nearby segment of DNA, and then perform a one-dimensional random walk (linear diffusion) on the nucleic acid until it found the operator. This mechanism would lead to a reduction in dimensionality: instead of a three-dimensional random walk, the repressor would only have to perform a one-dimensional walk. An alternative way of explaining the association rate is by postulating direct inter-segment transfer through looped intermediates. It is known that the Lac Repressor tetramer loops DNA by binding simultaneously to two segments (reviewed in Schleif, 1992), and this makes such a scenario plausible. Jeltsch *et al.* (1994) have proposed that the restriction endonuclease *Eco*RI physically tracks along DNA in order to find specific sites, since its ability to locate sites in a large piece of DNA could be compromised by the introduction of triple-stranded obstacles. Fickert and Müller-Hill (1992) have proposed that certain properties of the filter binding assay may have led to an overemphasis of the linear diffusion model.

In the years since Jacob and Monod's paradigmatic contributions, numerous gene-regulatory mechanisms, in addition to the repression of transcription initiation by direct

competition for binding between RNA polymerase and a repressor, have been described and analyzed in detail. Nonetheless, modulation of gene expression by proteins which specifically bind regulatory DNA sequences and in some way affect the efficiency of transcription initiation remains a central pillar of reductionist biology.

The Transcriptional Regulation of Gene Expression in Eukaryotes Unlike prokaryotes, eukaryotic cells have three RNA polymerases (Roeder and Rutter, 1969): polymerase I transcribes ribosomal RNA (rRNA), polymerase II transcribes messenger RNA (mRNA) and some small nuclear RNA (snRNA), and polymerase III transcribes 5S RNA and transfer RNA (tRNA). Each polymerase is a multi-polypeptide assembly of great complexity which nonetheless is incapable of accurately initiating transcription at the correct start site on its own. Rather, these polymerases require a number of accessory proteins known as general or basal transcription initiation factors (Sklar and Roeder, 1977; Parker and Roeder, 1977). Most of these basal factors are specific to each kind of polymerase, and are required for the transcription of most or all genes transcribed by a given polymerase. The promoters, that is, the minimal DNA sequence elements required for initiation of transcription, for the three polymerases are distinct, and it is the basal factors that are responsible for identifying them and recruiting the appropriate polymerase. The differential transcription of genes in various tissues, cell-types, and cell-cycle stages which underlies homeostasis and development is achieved by the interaction of the basal factors with other regulatory polypeptides, the gene-specific transcription factors, which bind to regulatory sequences present at a location pertinent to the genes they regulate and modulate the efficiency of transcriptional initiation (reviewed in Roeder, 1991; Zawel and Reinberg, 1993).

Basal and Gene-Specific Transcription Factors The basal factors assemble at DNA sequences present in the promoter, close to or at the starting point of transcription.

Two such elements have been described for RNA polymerase II promoters: the TATA-box (Gannon *et al.*, 1979) and the initiator (Smale and Baltimore, 1989). Transcription factor IID (TFIID), and specifically, its subunit polypeptide TATA-Binding Protein (TBP), recognizes the TATA-box, and initiates assembly of a multi-protein structure known as the pre-initiation complex (PIC) which allows accurate initiation of transcription. TFIID is also thought to be a major target of transcriptional activators. In a different and independent pathway, initiator element binding proteins such as YY1 (Seto *et al.*, 1991) and TFII-I (Roy *et al.*, 1991) can first bind to the initiator and then recruit the other required basal factors, possibly including TFIID (Roy *et al.*, 1993), and the polymerase in order to form the PIC. The only components of the PIC (basal factors or polymerase subunits) which are known to have sequence-specific DNA-binding activity are TFIID, YY1 and TFII-I, and their recognition of their respective DNA-sequence elements is a required first step for the orderly assembly of the PIC. Some RNA polymerase II promoters have either a TATA-box or an initiator, others have both.

Gene-specific transcription factors bind at DNA sequence elements close to or distant [up to many thousands of base pairs (kbp) away] from the promoter: these are traditionally referred to as promoter proximal elements and enhancers, respectively. It is instructive to return to the information content considerations introduced above. The *Escherichia coli* genome is approximately 3.9×10^6 bp. If a repressor is to find an operator present in a single copy in the genome and the repressor is capable of discriminating absolutely (*i.e.* it will bind to an operator with only one specified base in each position) then it will need an operator which is ~22 bp long. The Lac Repressor, which allows some "slop" in some positions, binds an operator which is ~30 bp long. If one considers the size of typical eukaryotic genomes (1×10^8 - 1×10^{11} bp), it becomes evident that the amount of information required uniquely to specify a binding site would imply binding sites hundreds of bp long. Such long binding sites would be very difficult to evolve and maintain. The

design of sequence-specific DNA binding proteins to bind such a long site with high specificity also poses many challenges. Finally, if a multicellular organism needed one transcription regulatory protein or factor per gene per developmental stage per tissue, its genome would be burdened with an astronomical number of regulatory proteins. More to the point, such long DNA-binding sites have not been found. The way this binding site length problem is circumvented by eukaryotic cells is to construct multi-partite binding sites, and to achieve cell- or tissue-specific gene regulation by employing a moderately large repertoire of transcription factors in a combinatorial manner. Each transcription factor then has limited specificity and affinity, but the total amount of information specified by the whole set of promoter proximal and enhancer elements controlling a gene is enough to achieve complex developmental and behavioral patterns. Typical binding sites for eukaryotic transcription factors span no more than ~10 conserved bp, and the dissociation constants of eukaryotic transcription factors for their recognition elements are commonly of the order of magnitude 10^{-9} to 10^{-10} M.

Activated Transcription Eukaryotic cells have most of their DNA packaged into chromatin by histones and other basic proteins. Transcriptional activation must achieve unpacking of the relevant genes in order to assemble the PIC at their promoters, efficient recruitment of polymerase to the initiation site, and efficient initiation of transcription. Apparently, transcriptional elongation is not drastically impeded by chromatin structure, but PIC formation is (reviewed, *e.g.* in Felsenfeld, 1992; Wolffe, 1994). Some transcription factors are thought to be able to bind to their recognition elements even if these are occluded by chromatin, destabilize the chromatin structure to make the promoters they control "visible" to the basal transcription machinery and other activators, and thus allow PIC assembly (see *e.g.* Wechsler *et al.*, 1994). Activators may also achieve increased rates of transcription of their genes by lowering the activation energy for any of a number of as yet ill-defined steps in PIC assembly and start of transcription proper by the polymerase (see

e.g., Klein and Struhl, 1994). Finally, some activators may achieve increased transcription of their target genes by increasing the life-time of the PIC in the promoters they regulate, if a given PIC may support multiple rounds of initiation by successive polymerase molecules. Activators are thought to be able to act at great distances from their promoters because they can come in contact with the PIC (either directly or through mediating proteins, "coactivators") by looping of DNA (reviewed, *e.g.* by Ptashne, 1988; Schleif, 1992) or other higher-order chromatin structure such as that which might be present at locus control regions (LCR, reviewed in Felsenfeld, 1992). Specific repression of genes, other than the overall repression due to chromatin structure, could in principle also be achieved by sequence-specific DNA binding proteins. The discussion of the structure of gene-regulatory proteins, the main subject of this Dissertation, will be start in the next section.

1.2 Structural Studies of DNA-Protein Interactions

Direct Readout and Other Simplistic Ideas Early speculation on how proteins might recognize specific sequences in double-helical DNA was based on the available structural knowledge of DNA and complexes of small "model compounds" with DNA, as well as on the chemical properties of amino acids and proteins. On the basis of the chemistry of the amino acids, hydrogen bonding and hydrophobic interactions between amino acid side-chains and the edges of DNA bases were thought to be likely candidates for interaction (the early literature is reviewed in Saenger, 1984; von Hippel and Berg, 1986). Sundralingam and Rao (1975) and Seeman *et al.* (1976) proposed on the basis of double helical DNA structure that a protein could directly "read" the sequence of a DNA molecule by recognizing the pattern of hydrogen bonding donors and acceptors present in the major groove of B-form DNA (making multiple hydrogen bonds with those atoms). The four possible base pairs would each present a unique pattern of donors and acceptors to the major groove. The minor groove was thought to be unsatisfactory for sequence-

specific recognition because the four base pairs do not present unique patterns of hydrogen-bond donors and acceptors. Concerning the protein structures which could interact with the major groove edges of nucleotide bases, both two-stranded anti-parallel β-sheets and α-helices were considered to be good candidates (Church *et al.*, 1977).

These early direct-readout propositions assumed that no sequence-dependent information useful for recognition was present in the local structure of the phosphodiester backbone, that torsional properties of DNA were not sequence dependent, that solvation and counter-ion condensation could not be exploited for sequence-specific recognition, and that the DNA (and the protein) underwent no significant deformation or "induced fit" upon association. The validity of these assumptions could be put to test for individual cases only when three-dimensional structures of a number of protein-DNA complexes were determined. In a fundamental epistemological sense, "direct readout" cannot be completely valid, because, strictly speaking, individual interatomic interactions can only be analyzed to yield interaction enthalpies, not free energies, and as such cannot allow us to understand fully the central issues: affinity and specificity (Mark and van Gunsteren, 1994). Although an argument can be made that electrostatic interactions can be analyzed to yield free energies (see *e.g.*, Gilson *et al.*, 1988), this assumes that the dielectric properties of concentrated solutions and the interior of macromolecules are known. This might be a trivial detail in other contexts, but in biology, where most meaningful interactions have free energies of only a few kilocalories per mole (see *e.g.*, Dill, 1990), these details are vitally important.

Families of DNA-Binding Proteins The isolation of a large number of sequence-specific DNA-binding proteins and the molecular cloning and sequencing of their genes which took place starting in the 1980's (and continues unabated now) had three great repercussions on our understanding of specific DNA-protein interactions. First, sequence alignments led to the realization that DNA-binding transcription factors, just as other

proteins, occur in families, that is, groups of proteins, presumably descendants of a common ancestor, that share structurally significant (Sander and Schneider, 1991) levels of sequence identity. Second, that for a number of families, the regions of conserved sequence did not span the entire lengths of the protein, suggesting the presence in these proteins of at least two domains, a DNA binding domain and some kind of transcriptional activation domain (reviewed, *inter alia* in Steitz, 1990; Harrison, 1991; Pabo and Sauer, 1992). This has subsequently been shown to be the case for many eukaryotic transcription factors; it is possible in most instances to splice the activation domain of one protein onto the DNA-binding of an entirely different transcription factor and to have the chimeric protein activate transcription through its binding to the recognition sequence corresponding to the DNA-binding domain. Third, knowledge of the sequence of these proteins made them available in the gram quantities needed for structural investigation through recombinant expression techniques. For the purposes of this discussion I will limit the use the word "domain" to refer to a polypeptide segment which can adopt a stable folded functional conformation on its own, and the word "motif" for conserved substructures within domains, incapable of autonomously acquiring a stable structure.

Structures of DNA-Protein Complexes In the ensuing years, a steadily growing number of structures of DNA-binding proteins or the DNA-binding domains of these proteins complexed with DNA have been determined, mostly by X-ray crystallography (Table 1). Historically, the first DNA-binding proteins whose structures were determined were the bacterial and bacteriophage proteins now classified as Helix-Turn-Helix (HTH) proteins. These proteins share a structural motif consisting of two α-helices separated by a three amino acid turn employed for interaction with, and recognition of the sequence of their target DNA, predominantly through the major groove. A member of this family, the Trp Repressor dimer in complex with the corepressor (tryptophan) and its operator DNA is illustrated in Fig. 1A. HTH family members whose structures have been determined in

Table 1. Structures of DNA-protein complexes[§] published before November 1994

Family	Subfamily	Name (source[#])	Resolution[*] (Å)
Helix-Turn-Helix (HTH)			
	Dimeric		
		Repressor (bacteriophage λ)	1.8
		Cro (bacteriophage λ)	3.9
		Repressor (bacteriophage 434)	3.2
		Cro (bacteriophage 434)	2.5
		Trp Repressor (*E. coli*)	1.9
		CAP (*E. coli*)	3.0
		Lac Repressor headpiece (*E. coli*)	(†)
	Monomeric		
		Engrailed homeodomain (*D. melanogaster*)	2.8
		Mat α2 homeodomain (*S. cereviseae*)	2.7
		Antennapedia homeodomain (*D. melanogaster*)	(†)
		Hepatocyte Nuclear Factor 3γ (HNF3γ, *R. norvegicus*)	2.5
	Tethered Multi-partite		
		Oct-1 POU domain (*H. sapiens*)	3.0
Zinc-Finger			
	TFIIIA		
		Zif268 (*M. musculus*)	2.1
		Gli (*H. sapiens*)	2.6
		Tramtrack (*D. melanogaster*)	2.8
	Nuclear Receptor		
		Glucocorticoid Receptor (GR, *R. norvegicus*)	2.9
		Estrogen Receptor (ER, *H. sapiens*)	2.4
	Binuclear Zn_2-Cys_6		
		Gal4 (*S. cereviseae*)	2.7
		PPR1 (*S. cereviseae*)	3.2
	GATA		
		GATA-1 (*G. gallus*)	(†)
	p53		
		p53 (*H. sapiens*)	2.2

E2		
	E2 (bovine papillomaviurs-1)	1.7
Leucine Zipper		
	GCN4 (*S. cereviseae*)	2.9
Helix-Loop-Helix (HLH)		
HLH-Zipper		
	Max (*H. sapiens*)	2.9
	Upstream Stimulatory Factor (USF, *H. sapiens*)	2.9
HLH		
	E47 (*H. sapiens*)	2.8
	MyoD (*M. musculus*)	2.8
β-Ribbon		
	MetJ Repressor (*E. coli*)	2.8
	Arc Repressor (bacteriophage P22)	2.6
TATA binding protein		
	TBP2 (*A. thaliana*)	1.9
	TBPc (*S. cereviseae*)	2.5

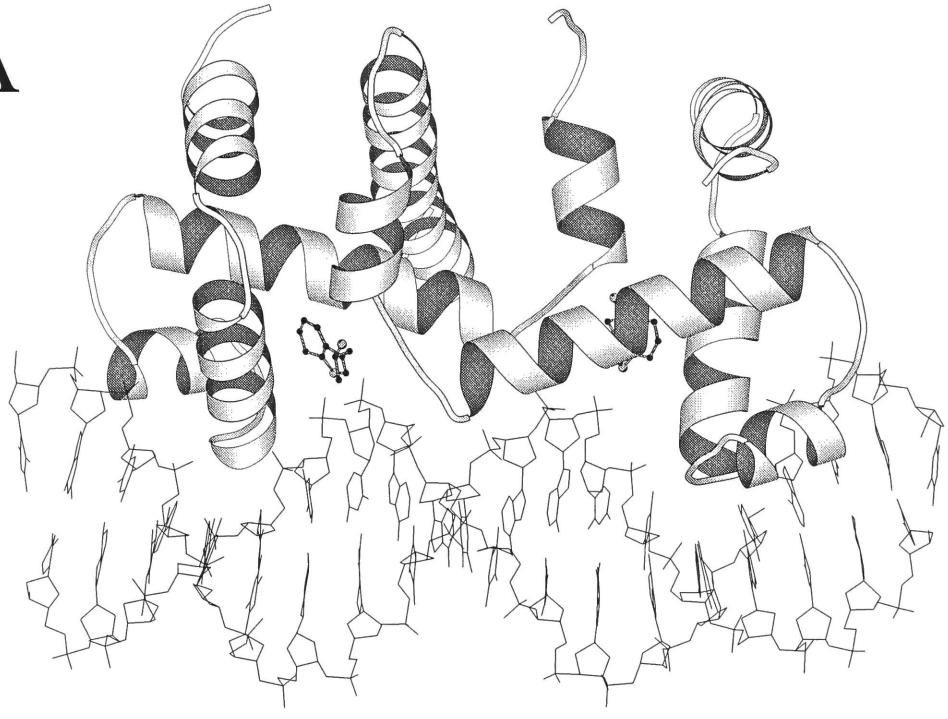
(§) Excludes enzymes; see references for details of protein construct employed. (#) Source of the sequence. (*) If several references are given, this is the highest resolution attained. (†) Determined by NMR spectroscopy. References: λ Repressor (Jordan and Pabo, 1988; Beamer and Pabo, 1992); λ Cro (Brennan *et al.*, 1990); 434 Repressor (Anderson *et al.*, 1987; Aggarwal *et al.*, 1988); 434 Cro, (Wolberger *et al.*, 1988; Mondragón and Harrison, 1991); Trp Repressor (Otwinowski *et al.*, 1988); CAP (Schultz *et al.*, 1991); Lac Repressor (Chuprina *et al.*, 1993); Engrailed (Kissinger *et al.*, 1990); Matα2 (Wolberger *et al.*, 1991); Antennapedia (Qian *et al.*, 1993a; Billeter *et al.*, 1993); HNF3γ (Clark *et al.*, 1993); Oct-1 (Klemm *et al.*, 1994); Zif268 (Pavletich and Pabo, 1991); Gli (Pavletich and Pabo, 1993); Tramtrack (Fairall *et al.*, 1993); GR (Luisi *et al.*, 1991); ER (Schwabe *et al.*, 1993); Gal4 (Marmorstein *et al.*, 1992); PPR1 (Marmorstein and Harrison, 1994); GATA-1 (Omichinski *et al.*, 1993); p53 (Cho *et al.*, 1994); E2 (Hegde *et al.*, 1992); GCN4 (Ellenberger *et al.*, 1992); Max (Ferré-D'Amaré *et al.*, 1993; this work); USF (Ferré-D'Amaré *et al.*, 1994; this work); E47 (Ellenberger *et al.*, 1994); MyoD (Ma *et al.*, 1994); MetJ (Somers and Phillips, 1992); Arc (Raumann *et al.*, 1994); TBP2 (Kim *et al.*, 1993a; Kim and Burley, 1994); TBPc (Kim *et al.*, 1993b); .

complex with DNA now include the dimeric Repressor and Cro proteins from bacteriophage λ and 434, the *E. coli* Lac Repressor headpiece, and Catabolite Activator Protein (CAP), the monomeric Engrailed, Matα2, and Antennapedia homeodomains, Hepatocyte Nuclear Factor 3γ, and the bipartite, tethered Oct-1 DNA binding domain. Other DNA-binding protein families have subsequently been characterized structurally (Table 1). These include a number of different protein architectures (five are listed in the Table) stabilized by the coordination of transition metals ("zinc fingers", Fig. 1B and 1C); proteins which employ a two-stranded anti-parallel β-sheet to bind to the major groove of their target DNA, the MetJ and Arc repressors; proteins which bind to the minor groove of their target sequence employing a ten-stranded anti-parallel β-sheet (TBP, Fig. 1D); and the Helix-Loop-Helix (HLH) proteins (distinctly different from the HTH proteins) whose structure I first determined as described in this Dissertation.

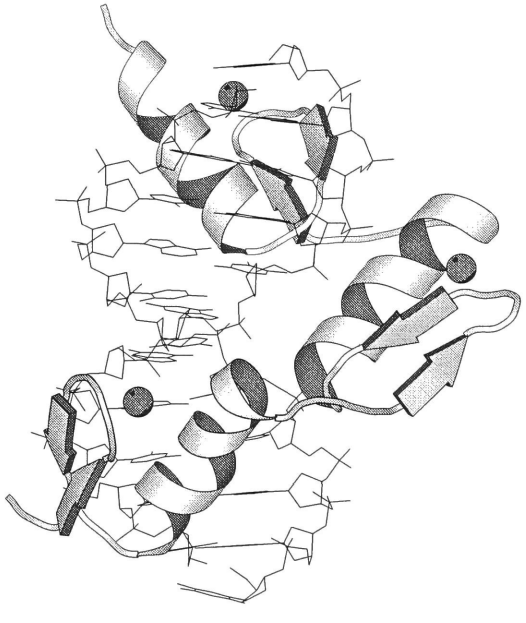
Induced Fit Determination of the first HTH-DNA structures at modest (~4 Å) resolution immediately revealed that both protein and DNA undergo conformational changes ranging from subtle to dramatic upon association. λ Cro protein dimer twisted by ~40° (relative to the structure of the free protein) and its operator B-form DNA was smoothly deformed by ~40° in the complex (Brennan *et al.*, 1990). CAP dimer was seen to bend DNA by ~90° by introducing two 40° kinks in the DNA (Schultz *et al.*, 1991). Determination of the cocrystal structure of the λ Repressor N-terminal domain bound to its operator at high resolution (I will restrict this term to mean better than 2Å resolution; for a discussion of resolution in X-ray crystallographic data see Swanson, 1988) confirmed biochemical and genetic evidence (Eliason *et al.*, 1985; Hurlburt and Yanofsky, 1992) that a N-terminal arm of the HTH domain, which appeared to be disordered in the free protein, became ordered in the DNA-protein complex and made minor groove contacts with DNA critical for sequence-specific recognition (Clarke *et al.*, 1991; Beamer and Pabo, 1992). This was in addition to apparently canonical direct readout-type contacts made in the major groove by the HTH

Figure 1. Representative examples of sequence-specific DNA binding proteins whose three-dimensional structures in complex with DNA have been determined. DNA is represented with thin lines, α-helical regions of the proteins as ribbons, β-strands as arrows, regions of irregular secondary structure as thin tubes, and bound zinc ions as spheres. Figures were prepared with the program MOLSCRIPT (Kraulis, 1991). **(A)** The dimeric *Escherichia coli* Trp Repressor (Otwinowski *et al.*, 1988; PDB accession number 1TRO), a canonical Helix-Turn-Helix (HTH) protein, bound to a duplex DNA containing its operator sequence. The co-repressor, tryptophan, is shown in ball-and-stick representation. **(B)** The canonical TFIIIA subfamily zinc-finger DNA-binding domain of the murine transcription factor Zif268 bound to its recognition site (Pavletich and Pabo, 1991; PDB accession number 1ZAA). **(C)** The canonical nuclear receptor subfamily zinc-finger dimer of the DNA-binding domain of the Glucocorticoid Receptor bound to its consensus site (Luisi *et al.*, 1991; PDB accession number 1GLU). **(D)** The TATA-Binding Protein isoform-2 from *Arabidopsis thaliana* bound to the adenovirus major late promoter TATA box (Kim *et al.*, 1993a); in this figure, the crystallographic DNA model has been augmented with 10 and 12 bp of B-form DNA in the left and right hand side, respectively. Atomic coordinates of TBP kindly supplied by J.L. Kim.

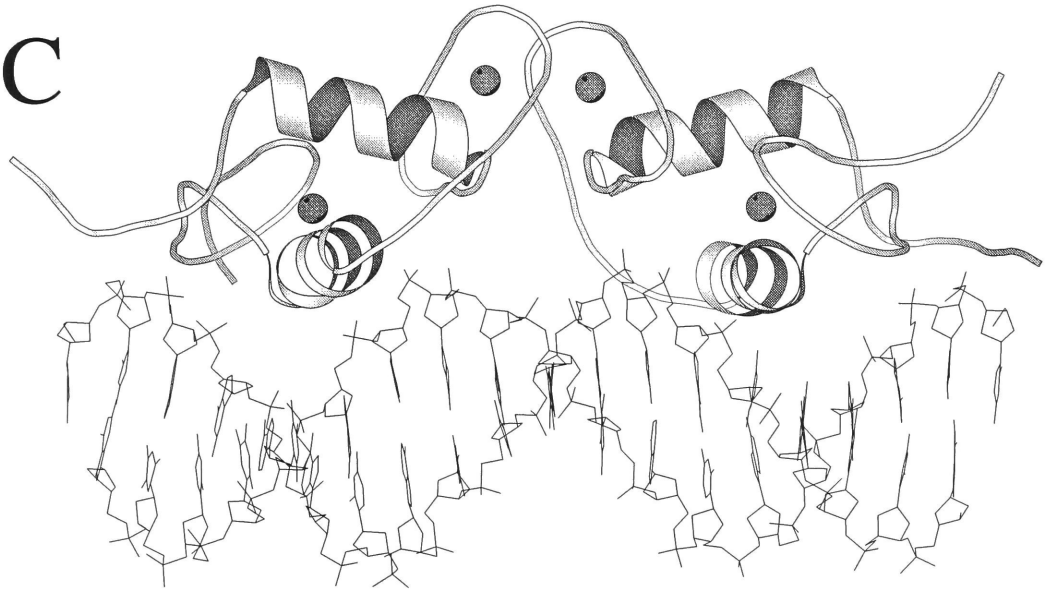
A



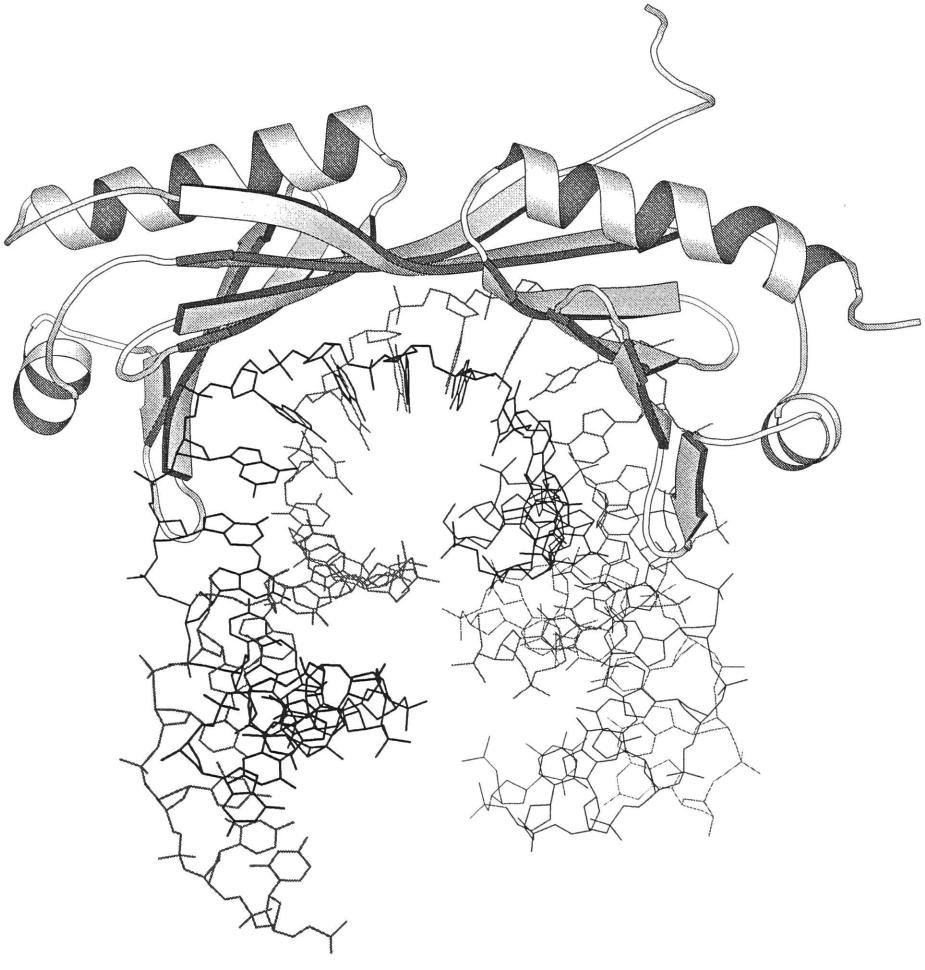
B



C



D



element. Subsequent work with the monomeric (but see Wilson *et al.*, 1993) eukaryotic HTH proteins, the homeodomain proteins, has shown the importance of similar N-terminal arm-minor groove edge contacts in recognition specificity (Qian *et al.*, 1993b; Billeter *et al.*, 1993; also see Clark *et al.*, 1993). Some of the most dramatic examples of the induced fit (Koshland Jr., 1958) that the protein moiety of DNA-protein complexes can undergo came when basic region-leucine zipper and basic region-helix-loop-helix proteins were investigated. These results, some of which I obtained in the course of the present work, will be further discussed in Chapter 4.

DNA Torsional Rigidity and Specificity The structure determination of the complex of the TATA-Binding Protein in complex with the TATA-box (Kim *et al.*, 1993a; Kim *et al.*, 1993b) reveled two unprecedented features which dealt further blows to the naïve direct-readout hypothesis. As can be seen in Figure 1D, this protein recognizes its target sequence (1) entirely through the minor groove, and (2) it does so by inducing a dramatic widening of the groove in order to expose the base edges to the flat bottom surface of the protein. If it was not obvious before, these structures made it evident that sequence dependent plasticity of DNA must be an important element of specificity (Hogan and Austin, 1987; Fujimoto and Schurr, 1990; early ideas on the importance of DNA plasticity are reviewed by Sigler, 1993). Classic measurements of DNA persistence length as a function of ionic strength (Harrington, 1978) had shown that DNA becomes a very flexible molecule as its negative charge is neutralized. This had led a number of authors to propose that DNA bound to positively charged proteins could be a rather flexible molecule (Manning, 1978). Sequence-dependent plasticity of DNA can have great biological importance.

Water The structure determination of the Trp Repressor-operator complex at high resolution (Otwinowski *et al.*, 1988; see also Lawson and Carey, 1993) revealed another

feature of DNA-protein recognition that had been ignored in the simplistic direct-readout view. This protein recognizes its target sequence entirely through indirect (through water) hydrogen bonding with the major groove edges of the DNA bases and contacts with the phosphodiester backbone. The importance of water in sequence specific recognition had been stressed by thermodynamic investigations such as those mentioned in Section 1.1, but ignored in the wake of the first structure determinations of DNA-protein complexes which were interpreted mostly in direct-readout terms. Four possible classes of DNA-protein contacts can be listed (1) direct protein-base contacts; (2) indirect (water mediated) protein-base contacts; (3) direct protein-backbone contacts; and (4) indirect (water mediated) protein-backbone contacts. The Trp Repressor-operator and subsequent structures which had well-ordered water molecules in the protein-DNA interface clearly displayed all four possible classes of DNA-protein contacts, and pointed out their importance for recognition. NMR spectroscopic data on the Antennapedia homeodomain-operator complex, interpreted in the light of molecular dynamics calculations, have started to give a time-dependent aspect to our picture of DNA-protein recognition, by showing that amino acid side-chains lying in the major groove of DNA which had been implicated biochemically to be important for sequence recognition, actually appear to explore multiple conformations in the micro- to nanosecond timescale, and water molecules seem to exchange rapidly between the protein-DNA interface and bulk solvent (Qian *et al.*, 1993b; Qian *et al.*, 1993a; Billeter *et al.*, 1993; the role of hydration in DNA recognition is reviewed by Sigler, 1993).

Dimerization and Cratic Entropy Many DNA-binding domains have been found to be dimeric (*e.g.* Trp Repressor, Fig. 1A; the Glucocorticoid Receptor DNA-binding domain, Fig. 1C). Biologically, this makes good sense for at least three reasons. First, oligomerization allows recognition of a relatively long sequence (lower uncertainty) with smaller proteins than would be required for a monomer recognizing the entire element. Second, oligomerization permits the incorporation of cooperative (or anti-cooperative)

responses into the recognition process. This often is coupled to binding concomitant induced-fit alluded to above. Third, oligomerization opens the way to combinatorial regulation of the biological response by forming homo- and hetero-oligomers which have different properties. The role of regulated dimerization and induced fit in biology have recently been discussed by Austin *et al.* (1994). From a chemical standpoint, however, dimerization is not free. The "cost of lost freedom", that is, the entropic losses due to reduction of translational and rotational degrees of freedom resulting from linking two molecules ("cratic entropy") has been estimated to be ~15 kcal mol^{-1} at ambient temperature (Finkelstein and Janin, 1989; Murphy *et al.*, 1994; see also Jencks, 1981). This cost must be paid for by the enthalpy of oligomerization, or more likely, since these reactions happen in water, by the entropy resulting from release of solvent molecules present in the interface of the protomers. This interface is, quite naturally, often hydrophobic. Clearly, the same thermodynamics apply to DNA-protein association, and we saw above that solvent/ion release is the major currency employed. Another way to pay for cratic entropy is with the formation of covalent bonds. This is exemplified by the TFIIIA subfamily of zinc finger proteins (such as Zif268, Fig. 1B) or the tethered multi-partite HTH proteins such as Oct-1, in which the independently folding subunits are part of the same polypeptide chain.

1.3 Helix-Loop-Helix Transcription Factors

Discovery The Helix-Loop-Helix family of transcription factors was first described by Baltimore and co-workers (Murre *et al.*, 1989a), who noticed the conservation of a 60 to 100 amino acid long sequence motif in a number of putative or *bona fide* transcription factors. The conserved amino acids were present in two blocks separated by a region of variable length and amino acid composition, which presumably would form a solvent exposed loop. Further, each conserved block presented a pattern of conserved hydrophobic amino acids and variable hydrophilic amino acids that suggested they might

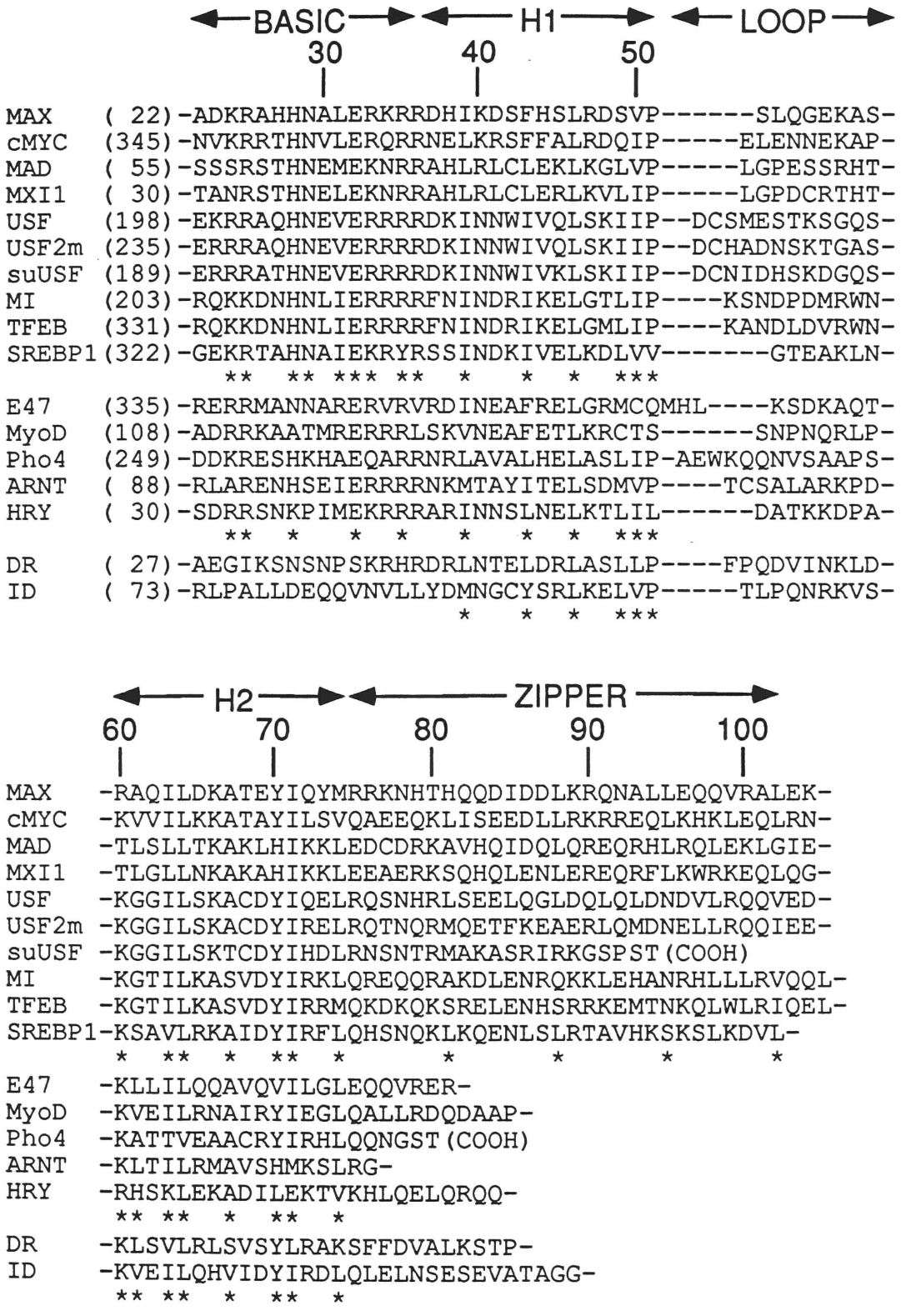
fold into amphipathic α-helices. The discovery came shortly after the proposal by (Landschultz *et al.*, 1988) that a different group of transcription factors (which were named "leucine zipper" proteins) shared a dimerization interface consisting of a coiled-coil (Crick, 1953b; Crick, 1953a) of two amphipathic α-helices, which presented a preponderance of leucine residues in the "d" position of the coiled-coil. In the leucine zipper proteins, the parallel coiled-coil dimerization interface is immediately C-terminal to a stretch of conserved, predominantly basic residues (Kouzarides and Ziff, 1988; Vinson *et al.*, 1989). This motif was consequently named the "basic region" and both together were proposed to form a DNA-recognition domain, the basic-leucine zipper or b/Z domain. An analogous stretch of basic residues (which shares no sequence similarity with the b/Z basic motif other than the presence of two clusters of basic amino acids separated by approximately seven residues) was promptly shown to be often present N-terminal to the HLH motif in many proteins (Prendergast and Ziff, 1989). These proteins thus became known as the basic-helix-loop-helix, or b/HLH proteins.

Biochemical experiments demonstrated soon thereafter that the b/HLH domain of a number of transcription factors was necessary and sufficient for dimerization and DNA binding. Dimerization was shown to be dependent on the HLH motif, while DNA binding appeared to involve predominantly the basic motif (Murre *et al.*, 1989b; Davis *et al.*, 1990; Voronova and Baltimore, 1990). (Strictly speaking, dimerization was demonstrated directly in none of these experiments, but was a parsimonious heuristic assumption. The evidence for oligomerization, formation of three complexes in EMSA experiments with proteins of two sizes, could be interpreted in terms of various other models.) In many instances, such as in the oncoproteins of the Myc family, the HLH was found to lie immediately C-terminal, and apparently in α-helical register, to a leucine zipper motif. The presence of what appeared to be two redundant dimerization interfaces, the HLH and the Z motifs, in these proteins was initially puzzling. Some helix-loop-helix proteins were also described that lacked the basic

region, but could form heterodimers with b/HLH proteins. These b/HLH-HLH heterodimers were inactive in DNA-binding, and hence heterodimerization resulted in repression of the transcriptional activity of the b/HLH partners (Benezra *et al.*, 1990). A partial sequence alignment of representative basic/helix-loop-helix/leucine zipper (b/HLH/Z), basic/helix-loop-helix (b/HLH), and helix-loop-helix (HLH) proteins, many of which play critical roles in homeostasis, the regulation development, and the control of cell proliferation, is shown in Fig. 2.

DNA-Binding Specificity The basic regions demonstrate considerable sequence conservation, and as would be expected if this motif constitutes the specificity determinant, the great majority of proteins belonging to the large helix-loop-helix family recognize a common CAnnTG DNA target, also known as the "E-box". In general, proteins with arginine at position 36 (in the Max numbering scheme) such as Max, Myc, USF, and Pho4 ("Class-B") bind to the palindrome caCGtg, while those binding to caGCtg, such as MyoD and E47 ("Class-A") have a hydrophobic residue at that position (Dang *et al.*, 1992). The modest size of the recognized sequence element prompted a number of researchers to use selection-amplification schemes with randomized DNA in the hope of defining extended binding sites. Blackwell *et al.* (1993) performed experiments with Myc/Max complexes as well as with a mutant MyoD protein that had its class A/B discriminating amino acid changed from the wild-type leucine to an arginine (as in Max). Solomon *et al.* (1993) performed experiments with Myc/Max as well as Max/Max complexes. Bendall and Molloy (1994) did their site selection experiments with USF. Cursory analysis of the selection results published by these authors showed that none of these proteins exhibited statistically significant (based on χ^2 tests) preference for particular bases outside the E-box, despite claims to the contrary in the papers. Max appeared to select with roughly equal frequency sites caCGtg and caTGtg (but never the palindrome caTAtg). The MyoD point mutation changed its specificity from class A to class B, as expected from the results of

Figure 2. Partial sequence alignment of representative basic/helix-loop-helix/leucine zipper (b/HLH/Z), basic/helix-loop-helix (b/HLH), helix-loop-helix proteins lacking a basic region (HLH). The consensus basic (b), helix 1 (H1), helix 2 (H2), and leucine zipper (Z) motif regions are indicated. This consensus is derived from both sequence alignments and structural investigations. Some positions with highly conserved residues are indicated with asterisks. For some of the illustrated sequences, the heptad repeat of the Z region may extend C-terminal beyond what is shown in the figure. Numbering at the top of the figure corresponds to the full-length human Max sequence. Numbers in parentheses preceding the basic region are the number of the first residue shown in this figure in the numbering scheme of the corresponding protein. Max (Blackwood and Eisenman, 1991; Prendergast *et al.*, 1991); cMyc (Murre *et al.*, 1989a); Mad (Ayer *et al.*, 1993); Mxi1 (Zervos *et al.*, 1993); human Upstream Stimulatory Factor ("43 kDa" USF, Gregor *et al.*, 1990); mouse USF2 ("44 kDa" USF, Sirito *et al.*, 1994; also see Blanar and Rutter, 1992); sea urchin USF (Kozlowski *et al.*, 1991); the product of the mouse *microphthalmia* locus (MI, Hodgkinson *et al.*, 1993); TFEB (Carr and Sharp, 1990); human Sterol Response Element Binding Protein-1 (SREBP-1, Yokoyama *et al.*, 1993); E47 (Murre *et al.*, 1989a); MyoD (Lassar *et al.*, 1989); Pho4 (Ogawa and Oshima, 1990); human Aryl hydrocarbon Receptor Nuclear Translocator (ARNT, Hoffman *et al.*, 1991); Hairy (Rushlow *et al.*, 1989); the human Dioxin Receptor (DR, Burbach *et al.*, 1992); human ID (Benezra *et al.*, 1990). The one letter amino acid code is: A, alanine; C, cysteine; D, aspartate; E, glutamate; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine, S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine.


BASIC
H1
LOOP
30
40
50
MAX (22)-ADKRAHHNALERKRRDHIKDSFHSLRDSVP------SLQGEKAS-
cMYC (345)-NVKRRTHNVLERQRRNELKRSFFALRDQIP-----ELENNEKAP-
MAD (55)-SSSRSTHNEMEKNRRAHLRLCLEKLKGLVP-----LGPESSRHT-
MXI1 (30)-TANRSTHNELEKNRRAHLRLCLERLKVLIP-----LGPDCRTHT-
USF (198)-EKRRAQHNEVERRRRDKINNWIVQLSKIIP--DCSMESTKSGQS-
USF2m (235)-ERRRAQHNEVERRRRDKINNWIVQLSKIIP--DCHADNSKTGAS-
suUSF (189)-ERRRATHNEVERRRRDKINNWIVKLSKIIP--DCNIDHSKDGQS-
MI (203)-RQKKDNHNLIERRRRFNINDRIKELGTLIP----KSNDPDMRWN-
TFEB (331)-RQKKDNHNLIERRRRFNINDRIKELGMLIP----KANDLDVRWN-
SREBP1 (322)-GEKRTAHNAIEKRYRSSINDKIVELKDLVV-------GTEAKLN-
E47 (335)-RERRMANNARERVRVRDINEAFRELGRMCQMHL----KSDKAQT-
MyoD (108)-ADRRKAATMRERRRLSKVNEAFETLKRCTS------SNPNQRLP-
Pho4 (249)-DDKRESHKHAEQARRNRLAVALHELASLIP-AEWKQQNVSAAPS-
ARNT (88)-RLARENHSEIERRRRNKMTAYITELSDMVP----TCSALARKPD-
HRY (30)-SDRRSNKPIMEKRRRARINNSLNELKTLIL------DATKKDPA-
DR (27)-AEGIKSNSNPSKRHRDRLNTELDRLASLLP----FPQDVINKLD-
ID (73)-RLPALLDEQQVNVLLYDMNGCYSRLKELVP-----TLPQNRKVS-
H2
ZIPPER
60
70
80
90
100
MAX -RAQILDKATEYIQYMRRKNHTHQQDIDDLKRQNALLEQQVRALEK-
cMYC -KVVILKKATAYILSVQAEEQKLISEEDLLRKRREQLKHKLEQLRN-
MAD -TLSLLTKAKLHIKKLEDCDRKAVHQIDQLQREQRHLRQLEKLGIE-
MXI1 -TLGLLNKAKAHIKKLEEAERKSQHQLENLEREQRFLKWRKEQLQG-
USF -KGGILSKACDYIQELRQSNHRLSEELQGLDQLQLDNDVLRQQVED-
USF2m -KGGILSKACDYIRELRQTNQRMQETFKEAERLQMDNELLRQQIEE-
suUSF -KGGILSKTCDYIHDLRNSNTRMAKASRIRKGSPST(COOH)
MI -KGTILKASVDYIRKLQREQQRAKDLENRQKKLEHANRHLLLRVQQL-
TFEB -KGTILKASVDYIRRMQKDKQKSRELENHSRRKEMTNKQLWLRIQEL-
SREBP1-KSAVLRKAIDYIRFLQHSNQKLKQENLSLRTAVHKSKSLKDVL-
E47 -KLLILQQAVQVILGLEQQVRER-
MyoD -KVEILRNAIRYIEGLQALLRDQDAAP-
Pho4 -KATTVEAACRYIRHLQQNGST(COOH)
ARNT -KLTILRMAVSHMKSLRG-
HRY -RHSKLEKADILEKTVKHLQELQRQQ-
DR -KLSVLRLSVSYLRAKSFFDVALKSTP-
ID -KVEILQHVIDYIRDLQLELNSESEVATAGG-

Dang *et al.* (1992). A site-selection experiment performed with E47 as well as the E47/MyoD heterodimer resulted in the asymmetric site caGGtg (Sun and Baltimore, 1991).

Some b/HLH/Z proteins have been found to be able to bind to sites distinct from their canonical Class-B E-box, and to exert biological effects from those. Both USF (Du *et al.*, 1993) and Myc/Max (Li *et al.*, 1994) have been shown to bind to and mediate transcriptional regulation (activation and repression, respectively) through the adenovirus major late promoter initiator element. Both appear to interact with the basal factor TFIID in effecting these regulatory functions. In addition, it has been reported that USF forms part of the multi-protein complex that assembles in the globin Locus Control Region (LCR), and in that context binds to the atypical E-box caCCtg (Bresnick and Felsenfeld, 1993). The Sterol Response Element Binding Proteins SREBP-1 and SREBP-2 bind and activate transcription of their target genes through the sterol response element (Yokoyama *et al.*, 1993; Hua *et al.*, 1993). These proteins also bind to Class-B E-boxes; it is not known whether they exert any regulatory function through the latter sites.

A small number of b/HLH proteins have been identified that posses divergent basic regions and bind to non-canonical E-boxes. *In vitro* site selection experiments with the protein Tal1 which has a divergent basic region of sequence -VVRRIFTNSRERWRQ- (starting at the same position as the aligned proteins of Fig. 2) but otherwise conforms closely to the HLH consensus, resulted in the target sequence caGAtg (Hsu *et al.*, 1994; Wadman *et al.*, 1994). The protein product of the yeast gene *RTG1*, which is involved in interorganelle communication between mitochondria, peroxisomes, and nucleus, has a basic region -PGSCGANFKNDRKRR-; its HLH region conforms closely to the consensus. Although its exact binding site has not been established, the restriction fragment that contains its biological target element (and to which the protein binds specifically *in vitro*) contains no canonical E-box (Liao and Butow, 1993). The Dioxin Receptor/ARNT (Fig. 2)

heterodimer appears to exert its biological function through binding to the atypical element ATGCGC (reviewed in Poellinger *et al.*, 1992). The *Drosophila melanogaster* negative regulator Hairy (Fig. 2) preferentially recognized the atypical site cACGCg in *in vitro* site selection experiments, and was shown to exert at least one of its developmental roles *in vivo* through such an element (Ohsako *et al.*, 1994).

Oligomerization Specificity: Networks of Transcription Factors Oligomerization appears to be more selective than DNA-binding, and only certain combinations of these proteins form biologically productive homo- and heterodimers. First, b/HLH/Z proteins do not associate with b/HLH proteins. Second, only limited subsets of either subfamily form productive oligomers. For example, the Myc oncoproteins, which are produced by the cell in response to proliferative signals and whose biological activities are predicated on DNA binding mediated by their b/HLH/Z domains, homodimerize weakly and do not bind DNA under physiologic conditions (reviewed in Amati *et al.*, 1993a). Conversely, the b/HLH/Z protein Max is expressed constitutively, homodimerizes and binds DNA efficiently, and represses transcription from promoters containing its target sequence (Blackwood and Eisenman, 1991; Prendergast *et al.*, 1991). When expression of the Myc proteins is induced, Myc and Max preferentially form heterodimers that compete for the same DNA sequences as the Max homodimer. Unlike Max, the Myc proteins contain an activation domain, enabling the Myc/Max heterodimers to stimulate transcription once bound to their cognate DNA (Amati *et al.*, 1993b; Mukherjee *et al.*, 1992). Recent characterizations of other b/HLH/Z proteins which oligomerize preferentially with Max, Mad (Ayer *et al.*, 1993) and Mxi1 (Zervos *et al.*, 1993) (Fig. 2) has led to the suggestion that Max plays a central role in orchestrating the biological activities of this group of b/HLH/Z transcription factors (Ayer and Eisenman, 1993). Neither Myc nor Max form heterodimers with other constitutively expressed b/HLH/Z or b/HLH proteins such as USF

or E47, which in turn have their own preferred dimerization partners (Kretzner *et al.*, 1992; Blackwood *et al.*, 1992; Prendergast *et al.*, 1992; Ayer and Eisenman, 1993).

A similar network of helix-loop-helix transcription factors plays a central role in the development of muscle cells. At least four myogenic b/HLH proteins, MyoD, Myogenin, Myf-5, and MRF4, four ubiquitous b/HLH proteins related to E12 and E47, and three negative regulators related to the HLH protein Id are expressed in these cells, and through their complex competitive interactions control their development (reviewed, *e.g.* in Sun and Baltimore, 1991; Shirakata *et al.*, 1993)

Upstream Stimulatory Factor USF is a widespread transcription factor that was first characterized as an activity from HeLa cell nuclei that bound to an upstream element of the adenovirus major late promoter (Sawadogo and Roeder, 1985; Carthew *et al.*, 1985; Miyamoto *et al.*, 1985), and stimulated transcription, possibly by direct interaction with the basal factor TFIID (Sawadogo and Roeder, 1985; Sawadogo, 1988; Meisterernst *et al.*, 1990; Workman *et al.*, 1990; Bungert *et al.*, 1992). Extensive purification of the HeLa nuclear activity yielded two polypeptides (Sawadogo *et al.*, 1988), the smaller of which was first cloned and sequenced revealing a protein of molecular weight 34 kDa with a highly conserved b/HLH/Z DNA-binding domain near its C-terminus (Gregor *et al.*, 1990). The recombinant protein expressed from this cDNA clone homo-oligomerized efficiently, bound DNA containing a caCGtg Class-B E-box motif with nanomolar affinity, and activated transcription in a manner indistinguishable from that of material purified from HeLa nuclear extracts (Pognonec and Roeder, 1991). USF has been shown to interact with a closely related protein called USF2 (in association with which it was purified from HeLa cell nuclei) (Sirito *et al.*, 1994), and with the initiator binding protein TFII-I (Roy *et al.*, 1991). Unexpectedly, when a mammalian expression library was screened for proteins which associated with the b/Z protein Fos, a protein which is virtually identical to USF2

was isolated (Blanar and Rutter, 1992). There are other reports in the literature of interactions between b/HLH and b/Z proteins (Bengal *et al.*, 1992). The biological relevance of these interactions between helix-loop-helix and b/Z proteins remains to be assessed.

1.4 Aims of this Work

Structure and Biological Functions of HLH Proteins The biological importance and relative functional simplicity of the helix-loop-helix family of proteins has made it the object of extensive genetic and biochemical investigation. Because the relatively small b/HLH/Z or b/HLH domains appeared to be both sufficient and necessary to mediate dimerization *and* DNA-binding, they also became promising systems for achieving understanding of biochemical activity as a function of three-dimensional structure. When I started my work on this system, lack of a three-dimensional structure was widely perceived to be a gaping hole in the understanding of the mode of action of these proteins. A three-dimensional structure would provide a conceptual scaffold with which to organize and interpret a large body of apparently disparate biochemical and genetic results addressing sequence specific DNA-binding as well as selective homo- and heterodimerization.

I therefore set out to determine the structures of some representative b/HLH/Z or b/HLH proteins in complex with DNA. The protein-DNA complexes and not the free proteins were the primary targets for structure determination because the complexes would potentially be informative on both functions mediated by these domains while the apo-proteins would only shed light on one function, namely oligomerization. An additional technical reason was that, as will be shown in Chapter 3, these proteins are partially unfolded in the absence of their target DNA, making successful crystallization unlikely.

The high degree of sequence conservation between the b/HLH/Z and the b/HLH subfamilies of proteins had led Murre *et al.* (1989) to predict that the structure of both subfamilies would be similar. I decided to concentrate my efforts on two b/HLH/Z proteins whose biological importance was well established: Max and USF. Both of these proteins had been shown to bind DNA containing canonical Class-B E-boxes, and exert their biological effects as homodimers, allowing me initially to avoid the more involved biochemistry of heterodimeric protein-DNA complexes. In addition, by working on proteins of the b/HLH/Z subfamily, I hoped to address the issue of the presence of two apparently redundant dimerization interfaces in these proteins. As is the case with most eukaryotic transcription factors, Max and USF contain both a phylogenetically conserved DNA-binding domain and divergent "effector" domains which are responsible for transcriptional modulation of their target genes. The structural basis of transcriptional activation or repression, I felt, was not yet ripe for direct investigation; I decided to focus on the conserved domain responsible specific DNA-binding and dimerization.

Experimental Strategy As is apparent from Table 1, X-ray crystallography is the technique of choice for the elucidation of the three-dimensional structure of protein-DNA complexes. X-ray crystallography requires preparation of large, well-ordered crystals of the molecule or molecular complex of interest; crystallization requires preparation of multi-milligram amounts of highly purified macromolecules. Therefore, a significant fraction of the experimental effort had to be devoted to biochemical purification. I decided to take advantage of the availability of large amounts of highly purified helix-loop-helix proteins and DNA to perform, in addition to structural determination, biochemical characterization and, once and for all, settle simple but fundamental issues such as stoichiometry, oligomerization state, conformational heterogeneity, and affinity. This Dissertation contains results from both experimental approaches, and employs them to present a coherent picture of the mode of action of b/HLH/Z and b/HLH proteins.

Chapter 2

Materials and Methods

Reagent purity was of great importance to the work reported here. Therefore, a significant amount of effort was devoted to achieving apparent homogeneity in macromolecular preparations. The methods employed for protein and DNA preparation are described in Sections 2.1 and 2.2, respectively. The fundamental assay employed to determine DNA-binding affinity and specificity, *i.e.* biochemical activity, of protein (and, incidentally, DNA) preparations was the electrophoretic mobility shift assay (EMSA) or gel-shift. The conditions under which this assay was carried out for the various proteins investigated in this work are summarized in Section 2.3 Conditions under which circular dichroism spectroscopy, dynamic light scattering (DLS), and analytical ultracentrifugation were performed are described in Sections 2.4, 2.5, and 2.6, respectively. The qualitative bimolecular ligation assay is described in Section 2.7. The general strategy employed for finding cocrystallization conditions for the various protein-DNA complexes investigated for

this work, and the exact conditions employed ultimately for growing data-collection quality crystals are detailed in Section 2.8. The crystallographic methodology that was employed is summarized in Sections 2.9 and 2.10. More details on structure determinations and refinements are given in Sections 3.2 and 3.4 of Chapter 3, Results. The use of DLS for assessing crystallizability of macromolecules and macromolecular complexes is discussed in the Appendix.

2.1 Protein Purification

General Considerations All proteins were expressed in soluble form in *Escherichia coli* BL21(DE3)pLysS, employing the T7 polymerase system (Studier *et al.*, 1990). The level of overexpression varied considerably from construct to construct: the yield of purified USF b/HLH was 50 milligrams per liter of culture; that of c/sHry was only 0.5 milligrams per liter of culture. All the helix-loop-helix proteins expressed were positively charged at neutral pH and bound tenaciously to bacterial DNA. This constituted the main hurdle for purification, since the proteins would not bind to ion-exchange chromatographic media when associated with the bacterial DNA. The bacterial DNA was stripped off the various USF constructs either by fractionation with ammonium sulfate or by exchanging the DNA for a molecular mimic: heparin, itself bonded to chromatographic medium. These strategies were ineffective for the other proteins. For those, DNA was eliminated by forming an insoluble complex with polyethyleneimine (PEI, Burgess, 1991) at low to moderate ionic strengths; the positively charged HLH proteins remained in solution together with excess PEI. The PEI precipitation also resulted in the incidental, and welcome, removal of a significant fraction of negatively charged *E. coli* proteins from the crude extracts. The excess PEI (which would bind irreversibly to cation-exchange media) was separated from the HLH proteins by precipitating these with ammonium sulfate. The redissolved HLH proteins could then be purified readily by ion-exchange and other

standard chromatographic techniques. The protease inhibitor "cocktail" was derived from that described by Pognonec *et al.* (1991). All mass spectrometric measurements were performed by S. Cohen and B. Chait (The Rockefeller University).

Upstream Stimulatory Factor (USF) Full-size (34.5 kDa, sequences of helix-loop-helix proteins prepared for this work are given in Fig. 3) recombinant human USF was expressed from the construct described by Kaulen *et al.* (1991). The product protein has a molecular weight of 33.5 kDa, but migrates with an anomalous apparent weight of 43 kDa on SDS polyacrylamide gels, and has been referred to in the literature as the "43 kDa form of USF". Cells were grown to an optical density 1.0 at 596 nm in M9ZB (Studier *et al.*, 1990), induced by the addition of IPTG to a final concentration of 0.5 mM for 5-8 h at 30°C. Cells were harvested by low-speed centrifugation and lysed by sonication in a buffer containing 500 mM NaCl, 10% (v/v) glycerol, 1 mM EDTA, 20 mM Tris-Cl pH 8.0, 1 mM PMSF, 1% (v/v) aprotinin (of the 15-30 trypsin inhibitor unit/ml solution distributed by Sigma Chemical Co.), 0.1% (v/v) NP-40, and 5mg/ml leupeptin. The clarified lysate was fractionated with ammonium sulfate and dialyzed against buffer A (100 mM KCl, 10% glycerol, 20 mM Hepes-KOH pH 7.5, 0.5 mM DTT, 0.5 mM PMSF, 0.5 mM EDTA, 0.1% (v/v) NP-40.) The dialysate was loaded onto a Mono-Q (Pharmacia) anion-exchange column pre-equilibrated with a buffer A containing 0.5% (w/v) octyl glucoside instead of NP-40. The NP-40 was fully washed away, and USF was eluted using a KCl gradient. USF eluted at 250 mM. The protein was dialyzed against buffer A, without any detergent, loaded onto a Mono-S (Pharmacia) cation-exchange column pre-equilibrated with buffer A, and a gradient identical to that used for the preceding Mono-Q chromatography was run; USF eluted at 330 mM KCl. The protein was estimated to be more then 98% pure by examination of serial dilutions with SDS-polyacrylamide gel electrophoresis (Laemmli, 1970) and Coomassie Blue staining. USF was stored in 10% glycerol, 1 mM PMSF, 100 mM KCl, 5 mM Hepes-KOH 7.5, 10 mM DTT, frozen at -70

Figure 3. Amino acid sequences in one letter code of proteins prepared for this study. Abbreviated names used throughout this work are given in parentheses. Amino acid residues derived from the expression vector are shown in lower case. The amino acid residues corresponding to the basic region consensus are in italics, those corresponding to the helix-loop-helix consensus are underlined. Serines which were mutated from cysteines are shown in bold letters. Numbering schemes corresponds to full-length proteins: human USF (Gregor *et al.*, 1990); human Max (Blackwood *et al.*, 1991; Prendergast *et al.*, 1991); Hairy (Rushlow *et al.*, 1989); SREBP-1 (Yokoyama *et al.*, 1993). The USF double C229S+C248S mutant, USF b/HLH/Z (C229+C248), and SREBP (C404) are not shown.

A Upstream Stimulatory Factor 1-310 (USF)

M KGQQKTAETE EGTVQIQEGA VATGEDPTSV AIASIQSAAT
FPDPNVKYVF RTENGGQVMY RVIQVSEGQL DGQTEGTGAI
SGYPATQSMT QAVIQGAFTS DDAVDTEGTA AETHYTYFPS
TAVGDGAGGT TSGSTAAVVT TQGSEALLGQ ATPPGTGQFF
VMMSPQEVLQ GGSQRSIAPR THPYSPKSEA PRTTRDEK*RR*
*AQHNEVERRR RD*KINNWIVQ LSKIIPDCSM ESTKSGQSKG
GILSKACDYI QELRQSNHRL SEELQGLDQL QLDNDVLRQQ
VEDLKNKNLL LRAQLRHHGL EVVIKNDSN

B Upstream Stimulatory Factor 1-260 (A/b/HLH)

M KGQQKTAETE EGTVQIQEGA VATGEDPTSV AIASIQSAAT
FPDPNVKYVF RTENGGQVMY RVIQVSEGQL DGQTEGTGAI
SGYPATQSMT QAVIQGAFTS DDAVDTEGTA AETHYTYFPS
TAVGDGAGGT TSGSTAAVVT TQGSEALLGQ ATPPGTGQFF
VMMSPQEVLQ GGSQRSIAPR THPYSPKSEA PRTTRDEK*RR*
*AQHNEVERRR RD*KINNWIVQ LSKIIPDCSM ESTKSGQSKG
GILSKACDYI QELRQSNHR

C Upstream Stimulatory Factor 197-310 C229S + C248S (b/HLH/Z')

m DEK*RRAQHNE VERRRRD*KIN NWIVQLSKII PDSSMESTKS
GQSKGGILSK ASDYIQELRQ SNHRLSEELQ GLDQLQLDND
VLRQQVEDLK NKNLLLRAQL RHHGLEVVIK NDSN

D Upstream Stimulatory Factor 197-260 C229S + C248S (b/HLH)

m DEK*RRAQHNE VERRRRD*KIN NWIVQLSKII PDSSMESTKS
GQSKGGILSK ASDYIQELRQ SNHR

E Max 3-113 H_6 (splice variant with nine amino acid deletion, residues 13-21, inclusive)

mgsshhhhhhssglvprgshmledp DNDDIEVESD AD*KRAHHNAL*
*ERKRRD*HIKD SFHSLRDSVP SLQGEKASRA QILDKATEYI
QYMRRKNHTH QQDIDDLKRQ NALLEQQVRA LEKARSSAQL QT

F Max 1-113 (splice variant with nine amino acid deletion, residues 13-21 inclusive)

M SDNDDIEVES DAD*KRAHHNA LERKRRD*HIK DSFHSLRDSV
PSLQGEKASR AQILDKATEY IQYMRRKNHT HQQDIDDLKR
QNALLEQQVR ALEKARSSAQ LQT

G Max 22-113

m AD*KRAHHNAL ERKRRD*HIKD SFHSLRDSVP SLQGEKASRA
QILDKATEYI QYMRRKNHTH QQDIDDLKRQ NALLEQQVRA
LEKARSSAQL QT

H Hairy 30-96 C50S (c/sHry)

m SD*RRSNKPIM EKRRRA*RINN SLNELKTLIL DATKKDPARH
SKLEKADILE KTVKHLQELQ RQQAAMQ

I Sterol Response Element Binding Protein-1 319-407 C404S (c/sSREBP)

mg QSRGE*KRTAH NAIEKRYRS*S INDKIIELKD LVVGTEAKLN
KSAVLRKAID YIRFLQHSNQ KLKQENLSLR TAVHKSKSLK
DLVSASGSG

°C, and quantified by UV spectrophotometry used an extinction coefficient of 1.48 x 10^4 M^{-1} cm^{-1}, calculated based on the amino acid composition of the protein (Edelhoch, 1967), at 280 nm. N-terminal Edman sequencing demonstrated complete cleavage of the initiation methionine. The double mutant C229S+C248S of USF (USF', Pognonec *et al.*, 1992) was expressed and purified in the same way, except that DTT was omitted from all solutions employed.

Upstream Stimulatory Factor 1-260 (A/b/HLH) The USF A/b/HLH construct was prepared in a manner analogous to the USF construct by P. Pognonec and R.G. Roeder (The Rockefeller University, unpublished). The clarified lysate (ammonium sulfate fractionation was omitted) of A/b/HLH overexpressing cells, prepared as with USF, was dialyzed overnight against buffer A (defined above for USF), loaded onto a Heparin-Sepharose (Pharmacia) column pre-equilibrated with the same buffer, and eluted at 300 mM KCl by running a linear gradient in the concentration of this electrolyte. The preparation was dialyzed against buffer A and further fractionated over a Mono-S (Pharmacia) column in the same manner as for USF. A/b/HLH eluted at 350 mM KCl. After dialysis against A, the protein was further purified by chromatography on a Mono-Q column (Pharmacia), as done for USF. A/b/HLH eluted at 130 mM, and was estimated to be more than 98% pure by gel electrophoresis as described for USF. A/b/HLH was dialyzed against the USF storage buffer, quantified by UV spectrophotometry using an extinction coefficient of 1.48 x 10^4 M^{-1} cm^{-1} at 280 nm, and frozen at -70 °C.

Upstream Stimulatory Factor 197-310 (b/HLH/Z and b/HLH/Z') The expression vectors for these proteins were also prepared by P. Pognonec and R.G. Roeder (unpublished). b/HLH/Z and b/HLH/Z' (the prime denotes the double mutation C229S+C248S) were purified in the same manner as A/b/HLH, except that the material eluted from the Mono-S column was judged to be more than 98% pure by gel

electrophoresis as described for USF, and no further purification steps were undertaken. Both proteins were dialyzed against the USF storage buffer, quantified by UV spectrophotometry using an extinction coefficient of 6.99 x 10^3 M^{-1} cm^{-1} at 280 nm, and frozen at -70° C. DTT was omitted during purification of b/HLH/Z'. MALDI mass spectrometry (Chait and Kent, 1992) of b/HLH/Z' demonstrated complete cleavage of the initiation methionine, and no detectable post-translational modifications or proteolysis (measured: 13424±2 a.m.u.; calculated: 13423 a.m.u.).

Upstream Stimulatory Factor 197-260 C229S+C248S (b/HLH) This construct was expressed in a manner similar to the previous proteins (P. Pognonec and R.G. Roeder, unpublished). Clarified lysate of b/HLH overexpressing cells was subjected to Heparin-Sepharose (Pharmacia) chromatography in the same manner as the above three proteins and eluted at 300 mM KCl. The protein, dialyzed against A without detergent, was subjected to subtractive chromatography over a Mono-Q (Pharmacia) column equilibrated with buffer A containing 0.5% octyl glucoside, and then loaded onto a Mono-S (Pharmacia) column equilibrated with the same buffer. A linear KCl gradient was run and b/HLH eluted at 200 mM KCl; it was estimated to be more then 99% pure by gel electrophoresis as described for USF. b/HLH was dialyzed against the USF storage buffer (minus DTT), quantified by UV spectrophotometry using an extinction coefficient of 6.99 x 10^3 M^{-1} cm^{-1} at 280 nm, and frozen at -70 °C. No DTT was included during the purification of this protein. MALDI mass spectrometry of b/HLH demonstrated complete retention of the initiation methionine, and no detectable post-translational modifications (measured: 7637±2 a.m.u.; calculated: 7638 a.m.u.).

Max 3-113 H6 The expression vector for this protein was constructed by ligating the MynΔCT open reading frame (Prendergast *et al.*, 1991) to a hexahistidine-tag sequence containing pET-3d (Novagen) expression vector (G. Prendergast and E.B. Ziff, New York

University Medical Center, unpublished). The sequence of the product protein is shown in Fig. 3. Transformed *E. coli* BL21(DE3)pLysS cells were grown in LB (Sambrook *et al.*, 1989) with 200 mg/l ampicillin and 50 μg/l chloramphenicol at 30 °C to an optical density of 1.0 at 596 nm; induction by addition of IPTG to a final concentration of 1 mM was allowed to proceed for five hours. Cells were collected by low-speed centrifugation, resuspended in a buffer containing 20 mM Hepes-KOH pH 7.5, 10% glycerol, 100 mM KCl, 5 mM $MgCl_2$, 1% (v/v) aprotinin, 5 mg/ml Leupeptin, 1 mM PMSF, 0.1% (v/v) NP-40 and 100 units/ml of DNase I (Boehringer-Mannheim, Grade II), and lysed by four cycles of freeze-thaw, followed by incubation with gentle stirring at 4°C for 20 minutes. The lysate was clarified by centrifugation and loaded onto a Chelating Sepharose Fast-Flow (Pharmacia) immobilized-metal affinity column charged with nickel (II) sulfate and equilibrated in 10% glycerol, 20 mM Hepes-KOH pH 8.4, 1M KCl, 150 mM imidazole, 1 mM PMSF. After washing the column for 20 column volumes with the same buffer, the protein was eluted by washing with a buffer containing 100 mM EDTA in addition. Neutralized polyethyleneimine (Burgess, 1991) was added to the resulting protein to a final concentration of 0.1% (w/v), and the suspension centrifuged. The supernatant was fractionated with ammonium sulfate to yield purified Max 3-113 H_6 in the 50% to 80% saturation cut. The protein was dialyzed into the same storage buffer employed for the above proteins (minus DTT) and quantified by UV spectrometry employing an extinction coefficient of 2.59 x 10^3 M^{-1} cm^{-1} at 280 nm. MALDI mass spectrometry confirmed complete cleavage of the initiation methionine and the absence of post-translational modifications or proteolytic cleavage (measured: 14,579±2 a.m.u.; calculated: 14,576 a.m.u.). The hexahistidine tag was cleaved by dialyzing the protein into a buffer containing 10% (v/v) glycerol, 20 mM Tris-HCl pH 8.4, 150 mM KCl, 2.5 mM $CaCl_2$, and incubating with 3 units of human thrombin (Calbiochem) per milligram of protein for 1 hour at 20 °C. The cleaved protein was purified by chromatography over a Mono-S column as described below for Max 22-113.

Max 1-113 The expression construct for this protein was also prepared by G. Prendergast and E.B. Ziff (unpublished). This protein was expressed and purified essentially as was Max 22-113 (below). It was quantified by UV spectrometry employing an extinction coefficient of 2.59 x 10^3 M^{-1} cm^{-1} at 280 nm. MALDI mass spectrometry indicated complete removal of the initiation methionine and lack of any detectable proteolysis or post-translational modification (measured: 12045±1 a.m.u.; calculated 12045 a.m.u.).

Max 22-113 The expression construct for this protein was also prepared by G. Prendergast and E.B. Ziff (unpublished.) Starter cultures of LB with 2 mg/ml of ampicillin were inoculated with glycerol stock and grown at 37°C to an optical density of 1.0 at 600 nm. Production cultures containing 200 mg/l of ampicillin were inoculated with starter culture and grown at 30°C to an optical density of 1.0 at 600 nm. Fresh ampicillin (2 mg/l) was then added, and cultures induced by the addition of IPTG to a final concentration of 0.5 mM. Cells were collected by low-speed centrifugation 4 hours after induction and resuspended in the same lysis buffer employed for Max 3-113 H6. The suspension was freeze-thawed three times, incubated for 40 minutes at 4°C and clarified by centrifugation. Neutralized polyethyleneimine was added to a final concentration of 0.5% (w/v) to the supernatant. After gentle mixing for a few minutes at 4°C, the suspension was again clarified by centrifugation, and then the supernatant was fractionated with ammonium sulfate. Max 22-113 was in the 50% to 80% saturation fraction. The pellet was dissolved in a buffer solution containing 25 mM KCl, 20 mM Hepes-KOH pH 7.5, 10% (v/v) glycerol, 0.5 mM EDTA, and 0.5 mM PMSF, and dialyzed against the same buffer. A Mono-S (Pharmacia) column was equilibrated in 25 mM KCl, 20 mM Hepes-KOH pH 7.5, 10% (v/v) glycerol, 0.5 mM EDTA, 0.5 mM PMSF, and 0.5% (w/v) octyl glucoside. The dialyzed sample was adsorbed onto the column and washed with this buffer until a

stable base-line was obtained. Protein elution was achieved by running a gradient of the equilibration buffer against a buffer of identical composition except for KCl which was 1M. Max 22-113 eluted at 0.5M KCl. The protein was dialyzed against the same storage buffer as the previous proteins and stored at -70°C. It was quantified by UV spectrometry employing an extinction coefficient of 2.59 x 10^3 M^{-1} cm^{-1} at 280 nm. MALDI mass spectrometry confirmed complete cleavage of the initiation methionine and the absence of post-translational modifications or proteolysis (measured: 10826±2 a.m.u.; calculated: 10826 a.m.u.).

Hairy 30-96 C50S (c/sHry) The expression plasmid for this construct was prepared by M. Caudy (Cornell University Medical School, unpublished) and was used to transform *E. coli* BL21(DE3)pLysS. The facile loss of this plasmid required an involved growth and induction scheme. Starter cultures were grown in NZCYM (Sambrook *et al.*, 1989) with 50 µg/ml carbenicillin and 34 µg/ml chloramphenicol at 30°C to an optical density of 0.6 at 596 nm. Fresh antibiotics (same amounts) were added to the starter cultures 30 minutes before using them to inoculate production cultures. These were in NZCYM with the same concentration of antibiotics, and were grown at 30°C to an optical density of 0.1 at 596 nm. Cells were then harvested by low-speed centrifugation, resuspended in fresh medium with the same amounts of antibiotics and grown at 30°C until they reached an optical density of 0.6 at 596 nm, whereupon they were induced by addition of IPTG to a final concentration of 1 mM. The endogenous RNA polymerase was inactivated 30 minutes after induction by adding rifampicin (freshly dissolved in methanol) to a concentration of 0.24 mM. Induction was allowed to proceed for a further 2.5 hours at 30°C, and then cells were harvested by low speed centrifugation, lysed and fractionated with PEI and ammonium sulfate in the same way as Max 22-113. The fractionated protein was dialyzed against 300 mM KCl, 10% glycerol, 20 mM Hepes-KOH pH 7.5, 0.5 mM EDTA, and 0.5 mM PMSF, and then loaded onto a SB-Spheron cation-exchange column (Integrated

Separation Systems) which had been equilibrated in the same buffer supplemented with 0.5% (w/v) octyl glucoside. c/sHry was eluted with a KCl gradient at approximately 600 mM. The protein was then dialyzed against a buffer composed of 1M KCl, 20 mM Hepes-KOH pH 7.5, 0.5 mM PMSF and 0.5 mM EDTA-NaOH. After dialysis, it was concentrated to 5 mg/ml by microfiltration (Centricon-3, Amicon) and loaded onto a 100 ml Superdex-75 Prep-grade (Pharmacia) gel-filtration column equilibrated and run in the same buffer used for dialysis. The resulting protein (≥ 98% pure, as judged by SDS-Tricine PAGE, Schägger and von Jagow, 1987) was dialyzed against 100 mM KCl, 10% (v/v) glycerol, 10 mM Hepes-KOH pH 7.5, quantitated by amino acid analysis (the protein does not absorb above 230 nm), and stored at -70°C. MALDI mass spectroscopy showed complete cleavage of the initiation methionine and no detectable proteolysis or post-translational modification (measured: 7892±2 a.m.u.; calculated 7890 a.m.u.).

Sterol Response Element Binding Protein-1 319-407 C404S (c/sSREBP) The expression plasmid for this construct was prepared in the laboratories of J. Goldstein and M. Brown (University of Texas Southwestern Medical Center, unpublished), and used to transform the same strain of *E. coli* employed for expression of the other constructs. It produces a protein encompassing residues 319 to 407 of the human SREBP-1 with the addition of a single glycine residue at the N-terminus (and the mutation C404S, Fig. 3). Starter cultures were grown at 30°C for 8h in LB to which had been added 50 μg/ml carbenicillin and 34 μg/ml chloramphenicol. Production cultures were grown in LB with 200 mg/L of ampicillin and 34 μg/ml chloramphenicol at 30°C until the optical density at 596 nm was 0.6, and induced by adding IPTG to a final concentration of 1 mM. Cells were collected after 3 hours of induction by low speed centrifugation and lysed by freeze-thaw as was done for Max 22-113. Before clarification, solid KCl was added to the lysates and dissolved to yield a final concentration of 1M. The clarified lysate was fractionated by addition of polyethyleneimine as was Max 22-113, and then fractionated with ammonium

sulfate. c/s SREBP precipitated around 50% saturation. The precipitate was dissolved in, and dialyzed against, a buffer composed of 300 mM KCl, 20 mM Hepes-KOH pH 7.5, 5% (v/v) glycerol, 0.5 mM PMSF and 0.5 mM EDTA. The dialyzed protein was loaded onto a SB-Spheron (Integrated Separation Systems) column equilibrated in a buffer of composition identical to the dialysis buffer but also containing 0.5% (w/v) octyl glucoside. The protein was eluted by running a gradient against a buffer of the same composition but containing 1M KCl. c/sSREBP eluted around 700 mM KCl. The protein was then dialyzed against a buffer composed of 1 M KCl, 20 mM Hepes-KOH pH 7.5, 0.5 mM PMSF and 0.5 mM EDTA. After dialysis and concentration to 10 mg/ml by ultrafiltration (Centricon-3, Amicon), the protein was loaded onto a pre-equilibrated 100 ml Superdex 75 prep-grade column (Pharmacia) from which it eluted with an elution volume of 64 ml. c/sSREBP was quantified by spectrophotometry employing an extinction coefficient of 2.6 x 10^3 M^{-1} cm^{-1} at 280 nm, dialyzed against 150 mM KCl, 10% glycerol and 10 mM Hepes-KOH pH 7.5 and stored at -70°C. MALDI mass spectrometry confirmed complete cleavage of the initiation methionine and the absence of post-translational modifications or proteolysis (measured: 10074±2 a.m.u.; calculated: 10074 a.m.u.).

Sterol Response Element Binding Protein-1 319-407 (SREBP) This protein, which is identical to c/sSREBP except for the presence of the cysteine at position 404), was purified in essentially the same way as was c/sSREBP, except that 10 mM DTT was included in every solution employed in the purification. It was also quantitated employing ultraviolet spectrometry and an extinction coefficient of 2.6 x 10^3 M^{-1} cm^{-1} at 280 nm. MALDI mass spectrometry confirmed complete cleavage of the initiation methionine and the absence of post-translational modifications or proteolysis (measured: 10090±2 a.m.u.; calculated: 10089 a.m.u.). The protein was fully oxidized employing the [copper (II) (1-10 phenanthroline)$_3$] complex (Martin *et al.*, 1993) as a catalyst. Equal volumes of stock solutions of 1-10 phenanthroline (450 mM in ethanol) and copper (II) sulfate (150 mM,

aqueous) were mixed to generate a light cyan solution of the complex. This solution was added to a solution of the protein at 30 μM in a buffer consisting of 100 mM KCl, 10% (v/v) glycerol and 10 mM Hepes-KOH pH 7.5 to a final copper ion concentration of 1.5 μM. The solution turned brown immediately (due to the reduction of the metal), and when incubated on ice for 20 minutes regained its initial light cyan color. Complete oxidation was confirmed by SDS-PAGE under non-reducing conditions.

2.2 DNA Purification

DNA Synthesis All DNA oligonucleotides prepared for this work were synthesized employing conventional phosphoramidite chemistry on an Applied Biosystems model 391 instrument. Most syntheses were performed at a nominal scale of 1.0 micromoles with all dimethoxytrityl protecting groups removed during synthesis. DNA was cleaved from the controlled pore glass column with ammonium hydroxide (30% aqueous, Aldrich) and deprotected at 65°C for 12 hours. DNA containing halogenated bases was cleaved in the same manner, but deprotected for 18 hours at 55°C employing a 3:1 (v/v) mixture of ammonium hydroxide (30%) and ethanol. This milder treatment improved the yield of halogenated DNA.

DNA Purification The optimized protocol involved three steps: the main purification was achieved by gel electrophoresis under denaturing conditions in preparative 8M urea polyacrylamide gels; the DNA was then concentrated and purified further by anion-exchange chromatography on a Mono-Q column (Pharmacia) at pH 10 and 55 °C; finally the DNA was desalted by reversed-phase chromatography on a C2 column with a mobile phase consisting of aqueous ammonium acetate and methanol. The anion-exchange step was performed in a manner similar to that described by Cubellis *et al.* (1985). Purity was initially monitored by phosphorylating a sample of purified DNA with ^{32}P phosphate and

analyzing the radioactive material by denaturing gel electrophoresis and autoradiography. Once the protocol was worked out, phosphorylation and gel electrophoretic analysis were performed only if problems either with synthesis or purification were suspected of occurring.

Halogenated DNA was purified in the same way, but was kept out of direct light at all stages, including gel-electrophoresis. The change in the pKa of 5-iodo or 5-bromo deoxyuridine upon (photolytic) loss of the halogen allowed resolution of halogenated DNA from degradation products by anion-exchange chromatography.

Quantitation and Annealing Single stranded DNA was quantitated by UV spectrometry employing extinction coefficients calculated to include nearest neighbor contributions (Puglisi and Tinoco Jr., 1989). For sequences likely to form stable secondary structures, the measurements were performed at 80-90 °C. For spectrometric quantitation of duplex DNA, the extinction coefficient was determined experimentally by measuring the absorbance of a solution of known concentration of duplex.

DNA was annealed by mixing stoichiometric amounts of the complementary strands, adding KCl to 100 mM and $MgCl_2$ to 1-5 mM, heating the solutions to 85°C for 30 minutes and then cooling to 4°C at a constant rate of 0.33 degree/minute. This was achieved conveniently by employing a Perkin-Elmer model 9600 thermal cycler. Typically, annealing was performed with DNA concentrations ranging from 0.1 to 3 mM, and the yield of annealed duplex, as estimated from autoradiography of non-denaturing polyacrylamide gel electrophoretic analysis, exceeded 98%.

2.3 Electrophoretic Mobility Shift Assays

USF and USF b/HLH/Z Single-stranded DNA was phosphorylated with γ-^{32}P ATP and T4 polynucleotide kinase, phenol/chloroform extracted, ethanol precipitated, purified over Sephadex G-25 (Pharmacia), and quantified by UV spectrophotometry. Binding reactions (10 µl each) contained 100 mM KCl, 10 mM Tris-HCl pH 8.0, 10% (v/v) glycerol, 5 mM DTT, 1 mM $MgCl_2$, 1 µM annealed labeled DNA, and USF. Binding reactions for b/HLH/Z', and b/HLH also contained 25 µg/ml BSA. DTT was omitted from reactions employing cysteine-free proteins. Binding reactions for USF were incubated on ice for one hour whereas binding reactions for the other constructs were incubated at room temperature for 30 minutes. After incubation, 5 µl of 20% (v/v) glycerol was added to each reaction, and the mixture was loaded onto a pre-run gel. USF was resolved on 4% acrylamide 39:1 (w/w) acrylamide/bisacrylamide gels cast and run in 380 mM glycine, 50 mM Tris base, 2.7 mM EDTA, at 10 V/cm for 1.5 hours at room temperature. b/HLH/Z, b/HLH/Z' and b/HLH binding reactions were resolved on 10% acrylamide 39:1 (w/w) Acrylamide/Bisacrylamide gels cast in the same buffer and run at 10V/cm for 2 hours at room temperature. Gels were dried and visualized by autoradiography on Kodak XAR film or on imaging plates. Imaging plates were scanned with a Molecular Dynamics Phosphorimager and bands were quantitated using the volume integration function on ImageQuant v.3.15 (Molecular Dynamics).

Max 3-113 H6, Max 1-113, and Max 22-113 The general procedure was as for USF and USF b/HLH/Z. Binding reactions (10 µl each) contained 100 mM KCl, 10% (v/v) glycerol, 10 mM Hepes-KOH pH 7.5, 1 mM $MgCl_2$, and 1 µM annealed double-stranded DNA, and were incubated at room temperature for 30 minutes. Complexes were resolved on pre-run 10% acrylamide 29:1 (w/w) acrylamide/bisacrylamide gels cast and run in 270 mM Tris, 222 mM boric acid, and 6.25 mM EDTA at 10 V/cm. Visualization and quantitation were as for USF.

c/sHry The general procedure was as for the previous proteins. Binding reactions (10 µl each) contained 100 mM KCl, 10% (v/v) glycerol, 10 mM Tris-HCl pH 8.0, 1 mM $MgCl_2$, 25 µg/ml BSA (Pierce), and 1 µM annealed double-stranded DNA, and were incubated at room temperature for 30 minutes. Complexes were resolved on pre-run 10% acrylamide 39:1 (w/w) acrylamide/bisacrylamide gels cast and run in 380 mM glycine, 50 mM Tris base, 2.7 mM EDTA, at 10 V/cm. Visualization and quantitation were as for USF.

SREBP The general procedure was as for the previous proteins. Binding reactions (10 µl each) contained 50 mM KCl, 5% (v/v) glycerol, 12.5 mM Hepes-KOH pH 7.5, 6 mM $MgCl_2$, and 1 µM annealed double-stranded DNA, and were incubated at room temperature for 30 minutes. Complexes were resolved on pre-run 8% acrylamide 39:1 (w/w) acrylamide/bisacrylamide gels cast, pre-run, and run in 540 mM Tris, 444 mM boric acid, and 13.5 mM EDTA at 10 V/cm. Visualization and quantitation were as for USF.

Synthetic Myc/Max and Max/Max dimers The general procedure was as for the above proteins. Synthetic proteins (Canne *et al.*, 1994) were purified by reversed-phase chromatography and lyophilized. They were dissolved in a buffer composed of 150 mM KCl, 20 mM Mes-KOH pH 5.5, 10% glycerol, and quantified by EMSA titrations against known concentrations of DNA. Stock solutions were frozen in liquid nitrogen and kept at -70°C. Incubations reactions for EMSA (10 µl each) contained 150 mM KCl, 10 mM Mes-KOH pH 5.5, 10% (v/v) glycerol, 1 mM $MgCl_2$, 0.5 or 1 mM annealed double-stranded DNA, and Myc/Max or Max/Max. Complexes were resolved in 6% acrylamide 29:1 acrylamide/bisacrylamide ratio gels cast, pre-run, and run in 540 mM Tris, 444 mM boric acid, and 13.5 mM EDTA at 10 V/cm. Visualization and quantitation were as for USF.

2.4 Circular Dichroism Spectroscopy

CD spectra were obtained with an AVIV model 62DS spectropolarimeter at room temperature, using 10, 1, 0.5, 0.1, or 0.01 mm path-length cylindrical quartz cells (Hellma), and 1.5 nm bandwidth, 1 or 10 second averaging time per point, measurements every 0.5 nm. 5-10 spectra were taken for each sample, averaged, blanked, and smoothed with a cubic spline algorithm employing software provided by the manufacturer. Blanks were taken before and after each data collection, and subtracted from each other to monitor instrument drift. Most spectra were collected with the proteins, DNA, or DNA-protein complexes in the same buffer used for the binding reactions for EMSA (without BSA), described above. Some spectra, specially those with dilute samples, were collected with the proteins or complexes in 100 mM Phosphoric acid-KOH pH 7.5, 100 mM KF. This buffer system allowed use of longer optical paths. The proteins and DNA-protein complexes were taken into the buffer used for CD by microdialysis, and the actual buffer against which samples were equilibrated was used as a blank, both for quantitation of the sample by UV spectrometry and for CD. Spectra were normalized on the basis of the concentration of amino acid residues present in the sample. Where indicated, DNA spectra were multiplied with the same normalization factor (adjusting for concentration) used to normalize the spectra of their protein-DNA complexes. Spectra were analyzed using the program CONTIN (Provencher and Glöckner, 1981).

2.5 Dynamic Light Scattering

DLS was performed with a model dp-801 DLS instrument (Protein Solutions, Inc.) which employs a 25 mW, 780 nm solid state laser, a ~ 7 µl quartz cell, and a avalanche photo-diode, detecting photons scattered at a fixed scattering angle of 90°. Apparent translational diffusion coefficients, molecular masses, hydrodynamic radii of gyration, and degree of sample polydispersity were calculated from the autocorrelation function using the

manufacturer's software. This software calculates the diffusion coefficient from the decay of the autocorrelation function by performing a non-linear least-squares fit of the autocorrelation coefficients to an exponential decay (the parameters fitted being the baseline, the gain and the decay constant). The translational diffusion coefficient is then the decay constant divided by the square of the amplitude of the scattering vector (Schmitz, 1990). The equivalent hydrodynamic radius of gyration of a hard sphere is computed using the Stokes-Einstein equation:

$$R_H = \frac{k_B T}{6\pi\eta D_T}$$

where R_H is the radius of gyration, k_B is Boltzmann's constant, T is the absolute temperature, η is the viscosity of the solvent, and D_T is the translational diffusion coefficient. The sample polydispersity is expressed as the standard deviation of the distribution of apparent hydrodynamic radii computed for a given sample. The apparent molecular mass (M_r) is calculated employing an empirical equation determined by the manufacturer

$$M_r = (1.549\, R_H)^{2.426}$$

which results from a regression analysis of the apparent hydrodynamic radii of gyration of globular proteins of known molecular mass and oligomerization state. Samples were taken into 100-150 mM KCl, 10 mM Hepes-KOH pH 7.5, and 10 mM (USF, USF b/HLH/Z, SREBP) or no (b/HLH/Z', b/HLH, Max1-113, Max 11-113, Max 22-113, c/sHry, c/sSREBP) DTT by microdialysis, and filtered through 200 Å filters (Anotop-10, Whatman) to remove dust particles. Sample concentrations ranging from 1 to 100 μM were chosen empirically to ensure that the dynamic light scattering signal exceeded the instrumental detection threshold. Twenty or more independent measurements were made from each sample, and the reported values are calculated arithmetic means. Plots of

apparent molecular weight distributions shown are histograms of the best-fit molecular weights calculated with data collected successively from same sample.

2.6 Analytical Ultracentrifugation

Analytical ultracentrifugation was performed in a Beckman model XL-A ultracentrifuge equipped with absorption optics and an An-60Ti rotor. Samples were dialyzed exhaustively against buffers free of glycerol, and the actual buffers employed for dialysis were used as blanks. Sedimentation equilibrium data were collected employing 6-channel 12 mm Epon centerpieces at wavelengths of 230-280 nm, performing radial scans with a step size of 0.002 cm repeatedly every 4 hours until change in the distribution of analyte was no longer detected. Data for analysis were collected with a radial step size of 0.001 cm and 50 individual scans were averaged to reduce noise. Data was analyzed using the non-linear least-squares minimization program ORIGIN (MicroCal Software, Inc.) fitting the experimentally measured radial distribution of absorption to the equation (Aune, 1978; Brenner *et al.*, 1993)

$$c_r = c_{m,r_0} \exp\left[\frac{(1-\bar{v}\rho)\omega^2}{2RT} M(r^2-r_0^2)\right] + \sum^{i} c_{m,r_0}^{n_i} K_{a_i} \exp\left[\frac{(1-\bar{v}\rho)\omega^2}{2RT} n_i M(r^2-r_0^2)\right] + E$$

where c_r and $c_{m,r0}$ are the absorptions of analyte at radius r and of the monomer at the top of the solvent column r_0, M is the monomer molecular mass, v is the partial specific volume of the analyte, ρ is the density of the solvent, K_{ai} is the association constant of the i-th oligomer which contains n_i subunits, ω is the angular velocity and E is the baseline absorption. Rotor speeds and temperatures are indicated in the figure captions. The sum is taken over all n species. Baseline absorption was estimated by "overspeeding" the sample at the end of the run and measuring the absorption in the solute depleted region of each cell. Partial specific volumes of the sample proteins were calculated from their amino

acid composition employing Cohn-Edsall amino acid partial specific volumes listed by Durchschlag (1986). Solvent density was measured using pycnometry.

2.7 Bimolecular Ligation

A DNA template with the sequence 5' cgttcggaat tcccagctct ccaagataat cccagactgc tctatggaga gacccaagtc tggccagagt aaagatggga ttctatccaa agcttggtat tatgttttta ggc**cacgtg**a ccggatccgc g 3' was synthesized and purified with a denaturing 8M urea gel after phosphorylation, and amplified by PCR using two 23 nucleotide primers complementary to either end of the corresponding strands (the E-box CACGTG is highlighted with bold letters). PCR was carried out under standard conditions, and ^{32}P was incorporated to a final specific activity of 3.5 x 10^{15} Bq/mol by including α-^{32}P dATP in the PCR reaction. The product was then cleaved using *Eco*RI restriction endonuclease (New England Biolabs) and, after Phenol/Chloroform extraction, all short sequences (PCR failure products and the short restriction fragments) were removed by gel purification under denaturing conditions. The purified DNA was quantified by Cerenkov counting, and annealed by linear cooling as with the EMSA probes. EMSA was employed to document efficient binding to b/HLH/Z through one specific site per DNA under conditions described above (not shown). Ligation reactions (10 µl each) all contained 1 µM annealed template, 50 mM Tris-HCl pH 7.8, 10 mM $MgCl_2$, 20 mM DTT, 1 mM ATP, 50 µg/ml BSA, and 0, 0.1, 1 or 10 units of DNA ligase (New England Biolabs) as defined by the manufacturer. Ligation reactions were started by adding the ligase (diluted in the same buffer) to the buffered DNA-protein mixture, incubated at 16 °C for 30 minutes, and stopped by adding Phenol/Chloroform and vortexing. The DNA was extracted, ethanol precipitated, and examined by electrophoresis on 10% acrylamide 8M urea denaturing gels. Kinased pBR322 DNA *Msp*I digest was used for molecular size markers. Visualization and quantitation were as described for the EMSA experiments.

2.8 Crystallization

General Screening Strategy The general strategy employed (1) screening of the protein constructs by DLS (see Appendix), (2) screening of crystallization conditions for a large number of variant DNA sequences for each selected protein, and (3) optimization of crystallization conditions for DNAs which yielded promising crystal forms. Most of the crystallization trials performed employed the hanging-drop vapor diffusion method (McPherson, 1990), although some experiments were performed with micro-dialysis (Zeelen and Wierenga, 1992) with 5-10 μl dialysis buttons (Cambridge Repetition Engineers). Microscopic seeding (Stura and Wilson, 1990) was routinely employed for either locating conditions which would support crystal growth but not nucleation, or for optimization of crystal size and morphology. Virtually all proteins employed for this study gave DNA complexes that were soluble in 100-150 mM supporting monovalent electrolyte. Therefore, "conventional" searches, where precipitant concentration was gradually increased during equilibration, were performed to seek crystallization conditions. SREBP gave insoluble complexes with a number of DNAs, and for these protein-DNA complexes, ammonium acetate vapor diffusion, where the concentration of this volatile salt was reduced gradually during equilibration (see *e.g.* Joachimiak and Sigler, 1991; Clark *et al.*, 1993; Kim *et al.*, 1993a) under a variety of conditions was employed. The most successful initial screening strategy for the majority of new protein-DNA complexes (this means both complexes of newly expressed and purified proteins as well as complexes of previously purified proteins with DNA of previously untried length or sequence) involved a pH *vs.* precipitant search at a fixed complex (0.5 - 1.0 mM) and supporting electrolyte (100-200 mM KCl) concentration. The buffers, precipitants, and divalent cations that were typically used for the first factorial search appear in Table 2.

Table 2. Typical parameters varied for first factorial screen of new DNA-HLH protein complexes

Buffers (all at *ca.* 100 mM)	Precipitants (concentration range)	Divalent cations (concentration range)
Sodium Acetate pH 3.5	MPD (20-30%)	$MgCl_2$ (0-200 mM)
Sodium Acetate pH 4.5	PEG 400 (20-30%)	$CaCl_2$ (0-200 mM)
Sodium Cacodylate pH 5.5	PEG 4000 (10-20%)	
Mes-KOH pH 6.5	PEG-MME 350 (20-30%)	
Hepes-KOH pH 7.5	PEG-MME 3500 (10-20%)	
Tris-HCl pH 8.5		
Bis-Tris propane pH 9.5		

Normally, the first screen included the buffers, the precipitants, and the divalent cations at one concentration each, *i.e.* 7 x 5 x 2 = 70 conditions. Because most HLH cocrystals appeared to grow preferentially at 4°C, this temperature was tried first.

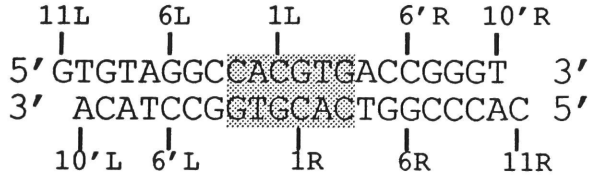
A rough factorial screen (Carter and Carter, 1979; Jancarik and Kim, 1991) of these variables could be performed with about 1 mg of protein-DNA complex. Temperature was either 4 or 20 °C. If promising crystals appeared, conditions were gradually fine-tuned by performing successively finer searches in the concentration of complex and precipitants, the pH, the concentration and chemical composition of the buffer, the concentration of mono- and divalent electrolytes, and the presence and concentration of additives such as metal ions, organic cations (spermine, spermidine), glycerol, ethylene glycol, etc. The complexes and conditions which eventually yielded crystals suitable for diffraction data collection are given below.

USF b/HLH+MLP21 $(b/HLH)_2$-DNA complex prepared by mixing two molar equivalents of b/HLH monomer with annealed, double-stranded DNA of the sequence depicted in Fig. 4A was concentrated by microfiltration to 1.4 mM (Centricon-3, Amicon) in a buffer containing 100 mM KCl and 10 mM Hepes-KOH pH 7.5. Cocrystals were grown at 4°C by vapor diffusion of hanging drops prepared by mixing equal volumes of complex and a reservoir solution consisting of 15% PEG 400, 15% glycerol, 100 mM KCl, 2.8 mM $MgCl_2$, 1.4 mM Cd(II) acetate, and 100 mM sodium acetate pH 4.75. Cocrystals grew over the course of several days to weeks to typical sizes of 0.3 x 0.3 x 0.2 mm^3.

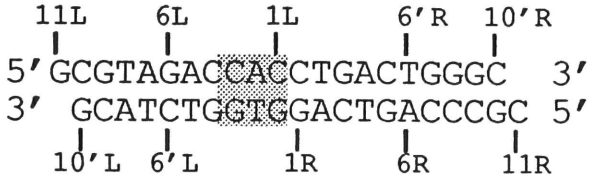
USF b/HLH+LCR21 These cocrystals were grown in essentially the same manner as USF b/HLH+MLP21, except the sequence of DNA employed was that shown in Fig. 4B and the reservoir solution was composed of 15% PEG 400, 5% (v/v) glycerol, 100 mM KCl, 100 mM sodium acetate pH 4.75, 2.5 mM $MgCl_2$, and 1.0 mM $BaCl_2$. Crystals grew within a few weeks to typical dimensions of 0.2 x 0.2 x 0.2 mm^3.

Figure 4. Sequences of double-stranded DNA oligonucleotides which yielded diffraction quality HLH cocrystals. (**A**) Adenovirus major late promoter 21-mer with G/C overhangs (MLP 21). Corresponds to positions -47 to -68 of the promoter (Ziff and Evans, 1978; Sawadogo and Roeder, 1985). (**B**) Human β-globin locus control region 21-mer with G/C overhangs (LCR 21) corresponding to positions 301 to 312 of the *Hin*dIII-*Xba*I fragment of Bresnick and Felsenfeld (1993) embedded in flanking sequences derived from MLP21. (**C**) Adenovirus major late promoter blunt-ended 22-mer, same positions as (A). (**F**) Sterol response element 20-mer (SRE 20) from the human low-density lipoprotein receptor gene, positions 38 to 57 of Gene Bank accession number I01624 (Yokoyama *et al.*, 1993). Sequences of double-stranded DNA oligonucleotides employed for most spectroscopic and EMSA studies reported herein: (**D**) specific 16-mer (abbreviated sDNA) derived from the adenovirus major late promoter. (**E**) non-specific 16-mer (abbreviated nsDNA) derived from D by introducing changes in positions 1R, 4R, and 6R. DNA sequences employed in crystallographic refinement of the Max (22-113)+MLP22 cocrystal structure: (**G**) double-stranded 11-mer employed in the first refinement. (**H**) one strand of the symmetric 22-mer employed in the final refinement. The positions of the crystallographic two-fold rotation axes are indicated in (G) and (H) by a lune. See section 3.2 for details. The E-box is shaded in all sequences (except on F where the sterol response element is shadowed).

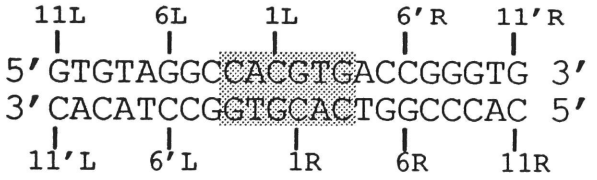
A


11L 6L 1L 6'R 10'R
5'GTGTAGGCCACGTGACCGGGT 3'
3' ACATCCGGTGCACTGGCCCAC 5'
10'L 6'L 1R 6R 11R

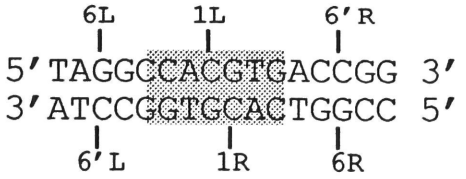
B


11L 6L 1L 6'R 10'R
5'GCGTAGACCACCTGACTGGGC 3'
3' GCATCTGGTGGACTGACCCGC 5'
10'L 6'L 1R 6R 11R

C


11L 6L 1L 6'R 11'R
5'GTGTAGGCCACGTGACCGGGTG 3'
3'CACATCCGGTGCACTGGCCCAC 5'
11'L 6'L 1R 6R 11R

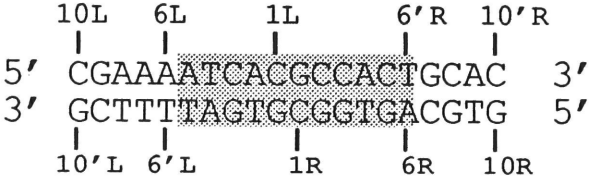
D


6L 1L 6'R
5'TAGGCCACGTGACCGG 3'
3'ATCCGGTGCACTGGCC 5'
6'L 1R 6R

E


6L 1L 6'R
5'AGGCCACCTGCCTGGG 3'
3'TCCGGTGGACGGACCC 5'
6'L 1R 6R

F


10L 6L 1L 6'R 10'R
5' CGAAAATCACGCCACTGCAC 3'
3' GCTTTTAGTGCGGTGACGTG 5'
10'L 6'L 1R 6R 10R

G


11 6 1
5'GTGTAGGTCAC 3'
3'CACATCCAGTG 5'
11' 6' 1

H


11L 6L 6R 11R
5'GTGTAGGTCACGTGACCTACAC 3'

Max 22-113+MLP22 (Max 22-113)$_2$-DNA complex prepared by mixing two molar equivalents of monomer with annealed, double-stranded DNA of the sequence shown in Fig. 4C was concentrated by microfiltration to 0.9 mM (Centricon-3, Amicon) in a buffer containing 100 mM KCl and 5 mM Hepes-KOH pH 7.5. Crystals were grown at 4 °C by microscopic seeding during vapor diffusion equilibration of sitting drops prepared by mixing equal volumes of complex and a reservoir solution consisting of 4.5-7.5% PEG 1000, 5-10% glycerol, 100 mM KCl, 2 mM $MgCl_2$, and 100 mM sodium cacodylate pH 5.5-5.75. Seeds were prepared from crystals of the same complex obtained at 12 °C by vapor diffusion of a mixture of the complex with an equal volume of a reservoir solution consisting of 100 mM sodium cacodylate pH 5.5, 5% (v/v) glycerol, 2 mM $MgCl_2$, 100 mM KCl and 3% (w/v) PEG 4000. Long (axial ratio ca. 1:10) hexagonal bars with a maximum diameter of 50 μm obtained under these conditions were crushed, resuspended in reservoir solution, and vortexed vigorously at 4°C. Serial dilutions of this seed stock were prepared up to a maximum dilution of 1:1,000,000, and stored at 4°C. The seed stocks remained viable for at least two months. Dilutions in the range 1:10,000 - 1:1,000,000 were employed for streak seeding with a cat's whisker. Hexagonal bars appeared within a few hours of seeding and grew over the course of several days to typical diameters of 0.4 mm and lengths of over 2 mm. Crystals of heavy atom derivatives were prepared in the same manner, using DNA in which 5-iodo dU had been substituted for T.

c/sSREBP+SRE20 The formation of these crystals appeared to be extremely sensitive to the exact protein:DNA ratio present in the crystallization drop. Thus, after formation of the complex by mixing 2.2 molar equivalents (presence of excess protein resulted in immediate precipitation of the complex) of annealed double-stranded DNA (of the sequence shown in Fig. 4F) with purified protein in 150 mM KCl, 10% glycerol, and 10 mM Hepes-KOH pH 7.5, and incubation at room temperature for 1-3 hours, the protein-DNA

complex was resolved from unbound DNA by gel-filtration chromatography on a 100 ml Superdex 75 Prep-Grade (Pharmacia) column equilibrated and run in 150 mM KCl, 25 mM Hepes-KOH pH 7.5, and 20 mM $MgCl_2$. The complex, double-stranded DNA, and single-stranded DNA had elution volumes of approximately 60, 80, and 93 ml, respectively, thus being easily resolved (peak-widths at the base-line were approximately 8 ml). The purified complex was concentrated by microfiltration to approximately 1 mM. Crystals were grown at 4°C by vapor diffusion of hanging or sitting drops prepared by mixing equal volumes of complex and a reservoir solution consisting of 18-25% MPD, 40-75 mM $MgCl_2$, 100 mM sodium acetate pH 4.5, 100-150 mM KCl. Crystals grew within a few days to typical dimensions of 0.2 x 0.2 x 0.05 mm^3.

2.9 X-ray Diffraction Data Collection

Data Collection with Laboratory X-ray Sources Data were collected in the laboratory with a Rigaku R-AXIS IIC automated imaging plate area detector using X-rays generated with a Rigaku RU-200 rotating anode source equipped with a copper anode, a 0.3 mm cathode, a graphite monochromator, and a double-pinhole 0.3 mm collimator. The generator was operated at 59 kV and 91 mA. Oscillation photographs were collected with a 100 μm raster. Cooling of crystals to temperatures between -20°C and +4°C was accomplished using a XR-85-1 Air Jet refrigeration unit (FTS systems) with a Molecular Structure Corporation controller. Cooling of crystals to near-liquid nitrogen temperatures was achieved employing a Molecular Structure Corporation cryogenic set-up.

Data Collection with Synchrotron X-ray Sources Diffraction data were also collected at beamline F-1 of the Cornell High Energy Synchrotron Source (CHESS F-1) in Ithaca, New York, and at beamline X-25 of the National Synchrotron Light Source (NSLS), Brookhaven National Laboratory, in Upton, New York. At both beamlines,

oscillation photographs were collected on Fuji imaging plates, scanned with a BAS-2000 (Fuji) scanner with a raster size of 100 μm. The focused, monochromated X-rays had wavelengths of 0.908 and 0.95 Å at CHESS and NSLS, respectively; the collimators employed had diameters of 0.1 and 0.2 mm, respectively. Crystals were cooled in the same way as in the laboratory.

Cryocrystallography For data collection at 4 to -20°C, crystals were mounted in capillaries at 4°C and cooled directly by insertion into the cold stream. For flash cooling of crystals to near-liquid nitrogen temperatures, crystals were first "cryoprotected" by transfer into a stabilizing solution, mounted in a loop (with a typical diameter of 0.2 - 0.4 mm, depending on crystal size) made with 10 μm opthalmic suture monofilament nylon (Ethilon 10-0, Ethicon Inc.), and flash-frozen either by insertion into a nitrogen gas stream at 100-110 K or by plunging into liquid propane held at liquid nitrogen temperature. The composition of the stabilizing solution for USF b/HLH+MLP21 and b/HLH+LCR21 was 15% PEG 400, 25% v/v glycerol, 100 mM NaCl, 100 mM sodium acetate pH 4.75, 5 mM $MgCl_2$, 1 mM cadmium (II) acetate, and 5% PEG 4000; and for c/sSREBP+SRE20 25% (v/v) MPD, 60 mM $MgCl_2$, 5% PEG 4000,100 mM sodium acetate pH 4.5, and 150 mM KCl.

Data Reduction Image files from both laboratory and synchrotron data collection were integrated and reduced using the programs DENZO and SCALEPACK (Z. Otwinowski, personal communication.) Unit cells were determined employing the R-AXIS IIc data processing package (Molecular Structure Corporation), manually employing DENZO, as well as by using the auto-indexing function of the data reduction program package HKL (Molecular Structure Corporation)

2.10 Structure Determination and Crystallographic Refinement

Max 22-113+MLP22 Native and derivative data-sets collected using the laboratory set-up were scaled using the ANSC program within the PROTSYS package (G.A. Petsko, personal communication). Weak and very strong isomorphous differences were filtered with the program ISOFIX of the same package, before the calculation of isomorphous differences. Patterson and difference Fourier syntheses were calculated using X-PLOR (Brünger, 1992). Derivative atom parameters were refined against both isomorphous and anomalous differences using HEAVY (Terwilliger and Eisenberg, 1983) in its PROTSYS implementation. Density modification was performed with SQUASH (Zhang and Main, 1990) and phase combination with COMBINE (Kabsch *et al.*, 1990). Model building was performed primarily using O (Jones *et al.*, 1991). Positional, simulated annealing (Brünger *et al.*, 1987), and *B*-factor refinement were all carried out using X-PLOR. The quality of the resulting atomic model was evaluated using PROCHECK (Laskowski *et al.*, 1993); DNA geometry was analyzed using the program CURVES (Lavery and Sklenar, 1989) and a global duplex axis. Solvent-accessible surface areas were calculated using the Lee and Richards (Lee and Richards, 1971) algorithm implemented in X-PLOR with a water probe radius of 1.4 Å. Figures were prepared using MOLSCRIPT (Kraulis, 1991), O, and QUANTA (Molecular Simulations, Inc.).

USF b/HLH+MLP21 Rotation search, Patterson correlation refinement, translation search, positional and *B*-factor refinement were all carried out using X-PLOR. The model was analyzed using the same programs employed for the first structure.

Chapter 3

Results

Section 3.1 describes the results of biochemical characterization of DNA binding by USF and its isolated DNA-binding domain. These studies set the stage for successful cocrystallization of a number of b/HLH proteins with DNA. A key methodological insight for successful cocrystallization, the use of dynamic light scattering (DLS) to assess crystallizability of macromolecules and macromolecular complexes, is described in the Appendix. The biochemical characterization of DNA binding by the helix-loop-helix protein Max which led to the successful cocrystallization of its b/HLH/Z DNA-binding domain with DNA, the characterization of those crystals, diffraction data collection from them, and structure determination are described in Section 3.2. The resulting structure is analyzed in some detail in Section 3.3. The structure determination and structure of the USF b/HLH DNA binding domain in complex with DNA is described and compared to that of the Max b/HLH/Z domain in complex with DNA in Section 3.4. Finally, work in

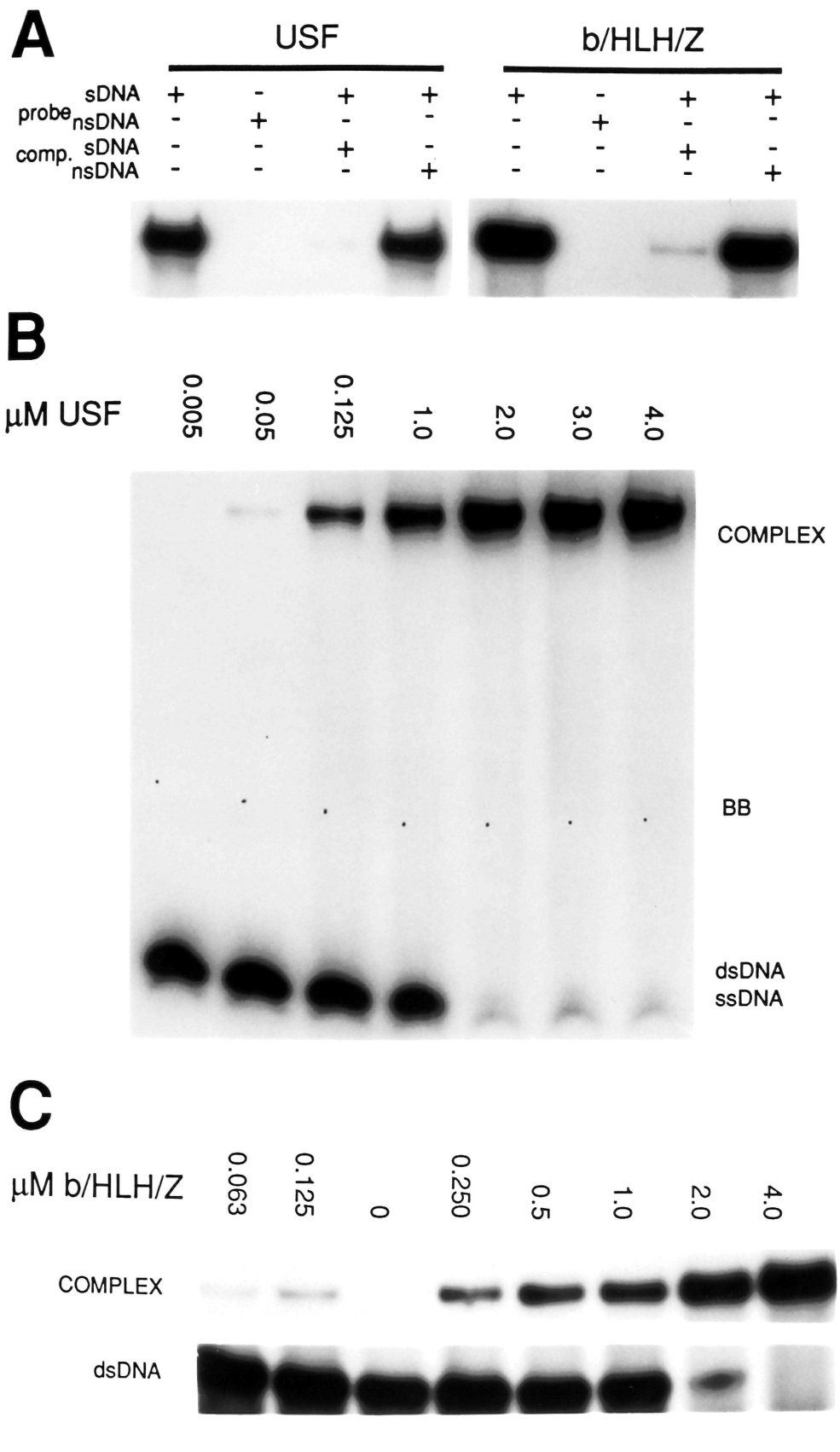
progress in various directions with a number of other helix-loop-helix proteins is summarized in Section 3.5.

3.1 Biochemical Characterization of DNA Binding by USF

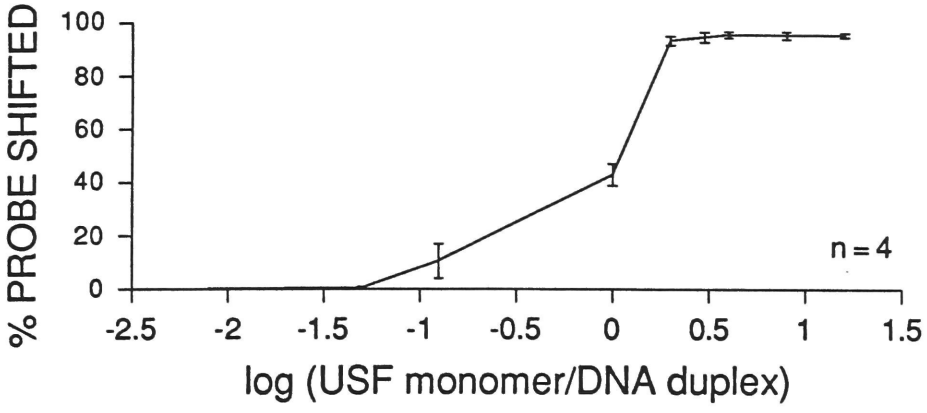
Recombinant USF and USF b/HLH/Z are Fully Active Full size USF, its intact DNA-binding domain (b/HLH/Z), and a construct of the DNA-binding domain missing the entire heptad repeat or leucine zipper element (b/HLH) were overexpressed and purified to apparent homogeneity. Recombinant USF binds specific DNA with an apparent dissociation constant of 1.3×10^{-9} M in electrophoretic mobility shift assays (EMSA) titration experiments (Pognonec and Roeder, 1991). EMSA demonstrate that both USF and b/HLH/Z efficiently bind 16 base pair (bp) oligonucleotides (sDNA, Fig. 4D), corresponding to the wild-type USF DNaseI footprint (Sawadogo and Roeder, 1985). In both cases, binding requires a protein to DNA duplex stoichiometry of 2:1, the apparent dissociation constants of b/HLH/Z and USF are comparable, and within experimental precision the purified, recombinant proteins have unit activity (Figs. 5A-5E). Neither USF nor b/HLH/Z form an electrophoretically stable complex with a mutant DNA (nsDNA, Fig. 4E) that incorporates mutations in only one half of the symmetric recognition element. In contrast, b/HLH does not form an electrophoretically-stable complex with specific DNA under conditions used for EMSA studies of both USF and b/HLH/Z. Under similar conditions, A/b/HLH forms a specifically retarded band containing ~5% of the labeled probe (data not shown).

Pognonec *et al.* (1992) had shown previously that recombinant USF which had both of its cysteines mutated to serines (C229S + C248S) retained wild-type activity, except that DTT could be omitted from binding reactions for EMSA experiments with the double mutant.

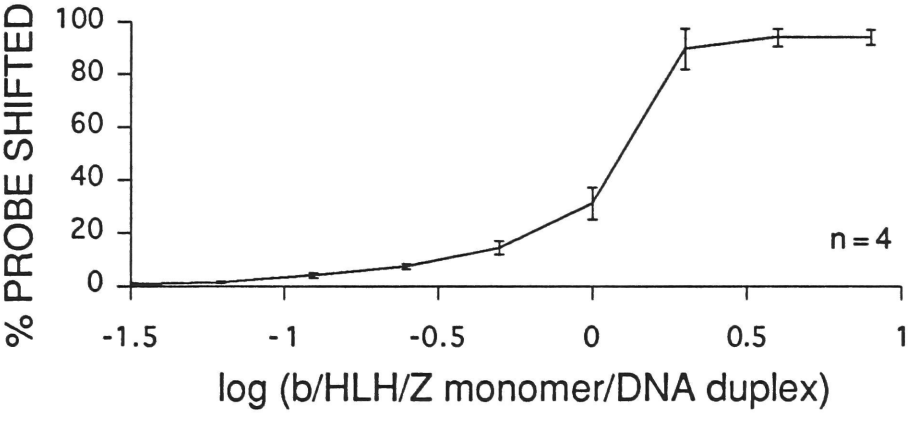
Figure 5. Determination of protein-DNA stoichiometries and DNA-binding activity and specificity for USF and USF b/HLH/Z. (**A**) Autoradiogram of representative EMSA experiments with USF and b/HLH/Z. ^{32}P-labeled DNA duplex and protein monomer concentration was fixed at 1 μM in each binding reaction. Cold competitor was added in a 100-fold molar excess where indicated. (**B**) Autoradiogram of a representative example of a titration with USF. Concentrations of USF monomer added to each binding reaction are indicated at the top of the figure. sDNA duplex concentration was 1 μM in each binding reaction. Migration positions of the protein-DNA complex, tracking dye (bromophenol blue, BB) double-stranded DNA (dsDNA) and single-stranded DNA (ssDNA) are indicated. There was no radioactivity detectable in this autoradiogram in the loading wells (not shown). (**C**) Sections of an autoradiogram equivalent to B but performed with b/HLH/Z. (**D**) Average quantitation of four independent experiments as shown in B. The ordinate is the percentage of the labeled DNA found in the shifted band in each lane; the abscissa is the Naperian logarithm of the molar ratio of protein monomer to duplex DNA. Bars represent standard errors of the mean. (**E**) Same as D but using b/HLH/Z instead of USF.


A
USF
b/HLH/Z
probe sDNA + - + + + - + +
nsDNA - + - - - + - -
comp. sDNA - - + - - - + -
nsDNA - - - + - - - +
B
μM USF
0.005
0.05
0.125
1.0
2.0
3.0
4.0
COMPLEX
BB
dsDNA
ssDNA
C
μM b/HLH/Z
0.063
0.125
0
0.250
0.5
1.0
2.0
4.0
COMPLEX
dsDNA

D


% PROBE SHIFTED
100
80
60
40
20
0
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
n = 4
log (USF monomer/DNA duplex)

E


% PROBE SHIFTED
100
80
60
40
20
0
-1.5
-1
-0.5
0
0.5
1
n = 4
log (b/HLH/Z monomer/DNA duplex)

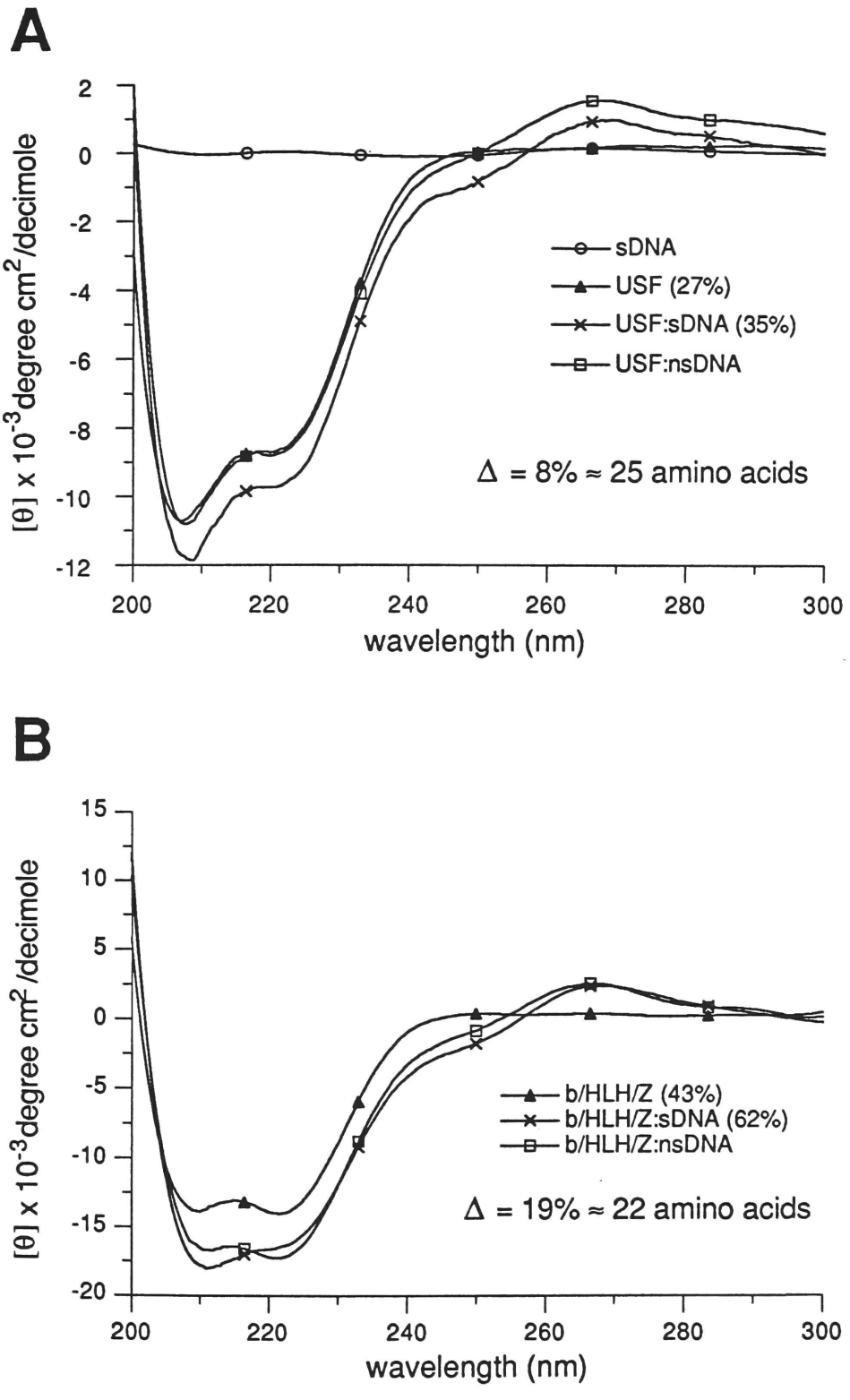
We performed EMSA experiments similar to those of Figs. 5 with full-size as well as b/HLH/Z constructs of USF incorporating these two mutations and also failed to see any deleterious effect (data not shown).

USF, b/HLH/Z, and b/HLH Undergo a DNA-Induced Folding Transition

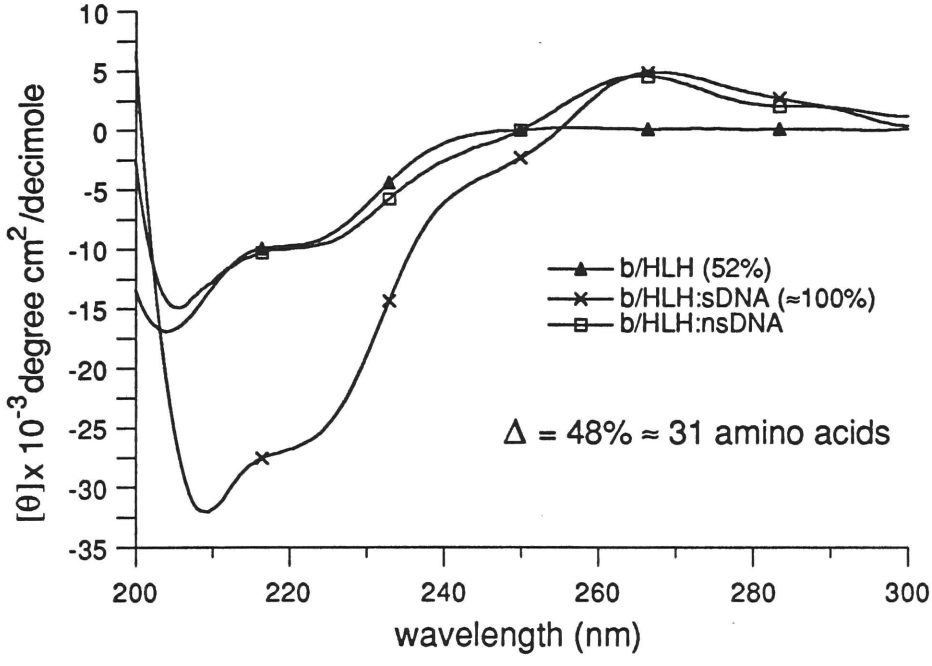
Interactions of the three proteins with sDNA and nsDNA were examined by circular dichroism (CD) spectroscopy. The spectra illustrated in Fig. 6A demonstrate that USF contains 27% α-helix in the absence of DNA. Incubation of USF with sDNA, but not nsDNA, provokes a negative deflection of the signal at 208 and 222 nm that is consistent with an increase in α-helix content of 8%, or about 25 amino acids per monomer. For reference, the CD spectrum of sDNA alone is illustrated in the same figure. b/HLH/Z undergoes a similar conformational change on DNA binding (Fig. 6B). Addition of sDNA results in an increase in α-helix content from 43% to 62%, a difference of 19% or 22 amino acids per monomer. Surprisingly, given the specificity exhibited in the EMSA experiment (Fig. 5A), this truncated form of USF appears to undergo a similar conformational change when mixed with nsDNA. In contrast, the minimal DNA-binding unit (b/HLH) undergoes an increase in α-helix content of 48%, or 31 amino acids, upon mixing with sDNA, but exhibits no conformational change when incubated with nsDNA (Fig. 6C).

The b/HLH spectra of Fig. 6C were obtained at a complex concentration of 13 μM in a buffer containing 10% glycerol. No change in the spectra occurred on increasing the concentration of the complex 10-fold in the same buffer, implying no further increase in α-helix content (data not shown). Interestingly, when the protein was incubated in a buffer containing potassium phosphate and potassium fluoride but no glycerol (a buffer system with minimal absorbance which allowed the use of long optical paths), this (and other, see Fig. 9B) helix-loop-helix proteins appeared to be less folded at the same concentrations.

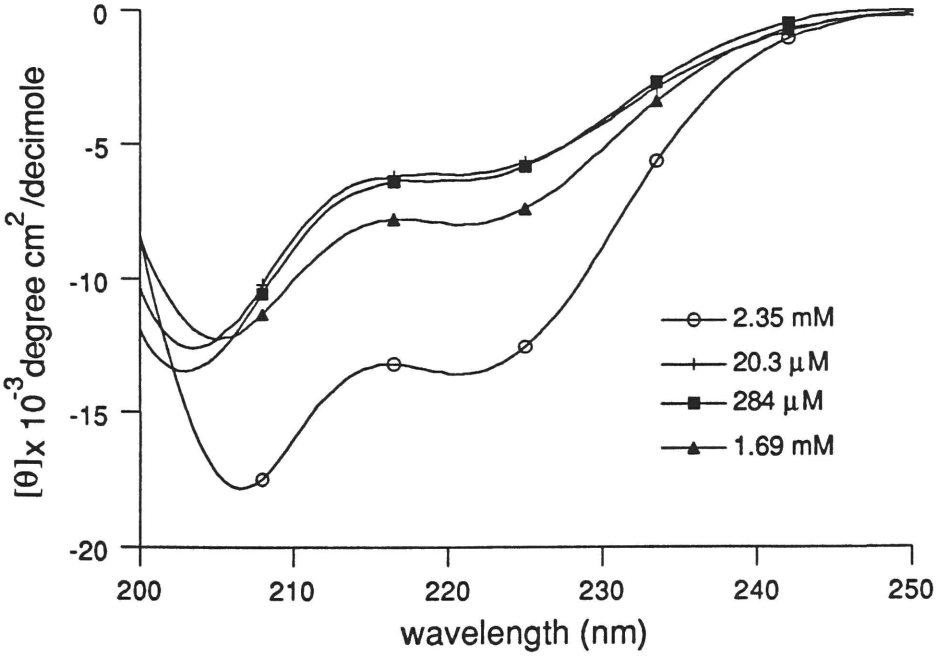
Figure 6. Normalized CD spectra of USF constructs. (**A**) Spectra of free USF, free sDNA, a 2:1 molar mixture of USF and sDNA (USF:sDNA) and a 2:1 molar mixture of USF and nsDNA (USF:nsDNA). The spectra were obtained with the macromolecular species at concentrations (on a monomer basis for the protein) close to 6 μM. The estimated α-helical contents of USF and USF:sDNA are shown in parentheses. The change in α-helical content indicated is per monomer. (**B**) Spectra of free b/HLH/Z, a 2:1 molar mixture of b/HLH/Z and sDNA (b/HLH/Z:sDNA), and a 2:1 molar mixture of b/HLH/Z and nsDNA (b/HLH/Z:nsDNA). The concentration of b/HLH/Z was close to 13 μM on a monomer basis for each experiment. The estimated helical contents of b/HLH/Z and b/HLH/Z:sDNA are indicated in parentheses. (**C**) Spectra of free b/HLH, a 2:1 molar mixture of b/HLH and sDNA (b/HLH:sDNA), and a 2:1 molar mixture of b/HLH and nsDNA (b/HLH:nsDNA). The concentration of b/HLH was close to 13 μM, on a monomer basis, for each experiment. The estimated α-helical contents of b/HLH and b/HLH:sDNA are indicated in parentheses. (**D**) Spectra of b/HLH obtained at the indicated concentrations of monomer. The spectra in this figure (D) were obtained with the protein in the phosphate/fluoride buffer system described in Section 2.4.


A
2
0
-2
-4
-6
-8
-10
-12
[θ] x 10-3degree cm2/decimole
200
220
240
260
280
300
wavelength (nm)
sDNA
USF (27%)
USF:sDNA (35%)
USF:nsDNA
Δ = 8% ≈ 25 amino acids
B
15
10
5
0
-5
-10
-15
-20
[θ] x 10-3degree cm2/decimole
200
220
240
260
280
300
wavelength (nm)
b/HLH/Z (43%)
b/HLH/Z:sDNA (62%)
b/HLH/Z:nsDNA
Δ = 19% ≈ 22 amino acids

C


10
5
0
-5
-10
-15
-20
-25
-30
-35
200
220
240
260
280
300
wavelength (nm)
[θ]x 10^-3 degree cm^2 /decimole
b/HLH (52%)
b/HLH:sDNA (≈100%)
b/HLH:nsDNA
Δ = 48% ≈ 31 amino acids

D

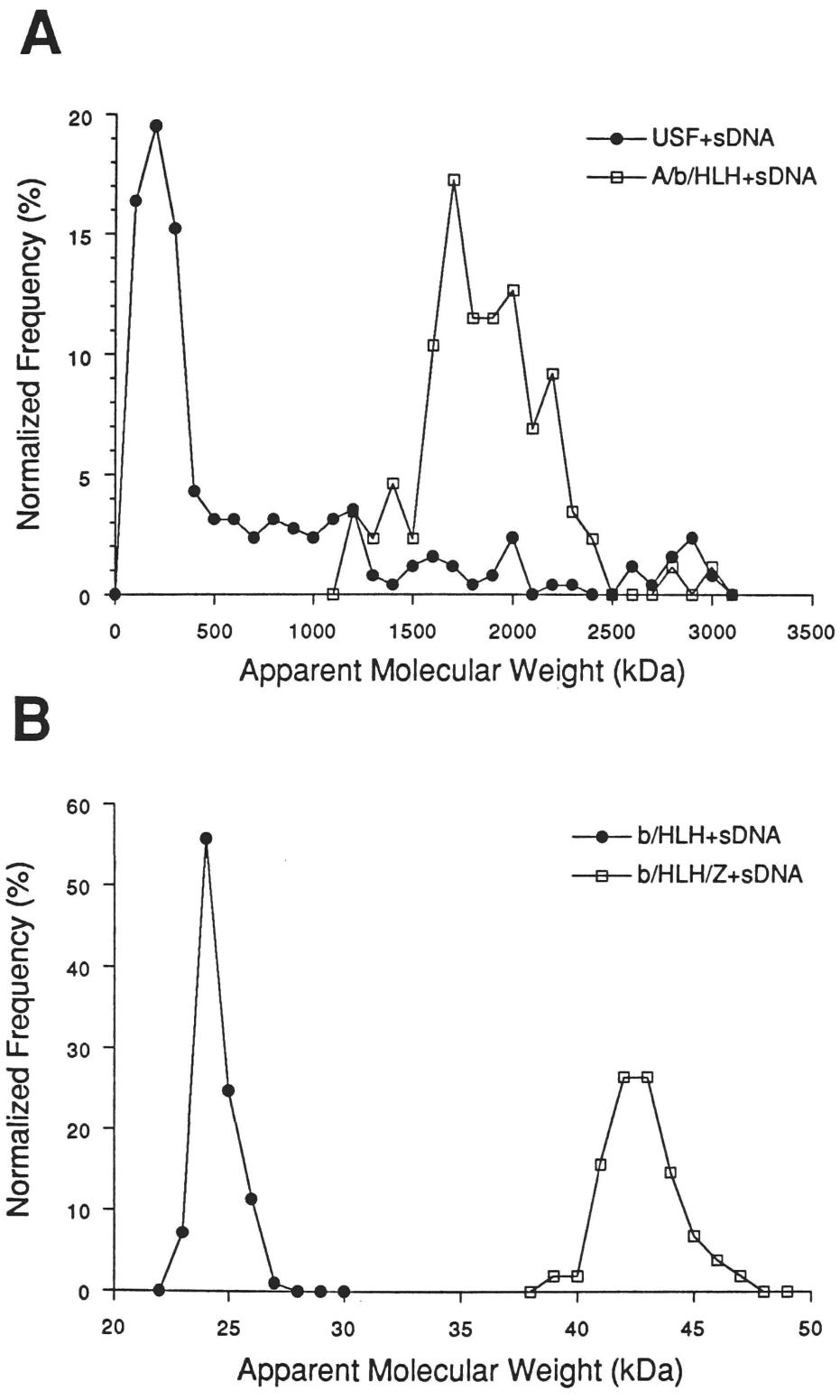

0
-5
-10
-15
-20
200
210
220
230
240
250
wavelength (nm)
[θ]x 10^-3 degree cm^2 /decimole
2.35 mM
20.3 μM
284 μM
1.69 mM

Spectra shown in Fig. 6D were collected with the USF b/HLH construct in this low-absorbance buffer, and it can be seen that the protein undergoes a concentration-dependent folding which does not become saturated (compare with Fig. 6C, b/HLH alone) up to a concentration of 1.7 mM.

The 2:1 protein monomer:DNA stoichiometry and the two-fold symmetry present in the recognition element suggest that a dimer binds symmetrically to it. This presumption is reinforced by the observed high cooperativity: mutations in one half of the binding element appear to abolish specific binding. The determination of the three-dimensional structures of helix-loop-helix proteins confirmed this.

Oligomerization States of USF, b/HLH/Z' and b/HLH Under a variety of solvent conditions, purified, recombinant USF exists in a highly aggregated and polydisperse state with an average oligomer mass exceeding 1 million daltons (Fig. 7A). Removal of the leucine zipper to yield A/b/HLH worsens the observed aggregation. In contrast, removal of the activation domain (here defined operationally as all amino acids N-terminal to the basic region, but see also Kirschbaum *et al.*, 1992) to give b/HLH/Z and b/HLH eliminates high order aggregation, and both smaller proteins are monodisperse in solution (Fig. 7A). The molecular weights measured by DLS for free and DNA-bound b/HLH and b/HLH/Z (at complex concentrations where the full folding transition is observed by CD spectroscopy) are given in Table 3. The macromolecular mass of b/HLH bound to sDNA predictably corresponds to a protein dimer complexed with one DNA duplex. However, the presence of the leucine zipper alters the behavior of b/HLH/Z dramatically. Under various solvent conditions, the complex of b/HLH/Z with sDNA occurs as a complex of four polypeptide chains and two DNA duplexes. In the absence of sDNA, b/HLH and b/HLH/Z exhibit masses consistent with those of protein dimer and tetramer, respectively.

Figure 7. Hydrodynamic characterization of some USF constructs. **(A)** Results of dynamic light scattering measurements (DLS) on full-size USF+sDNA, A/b/HLH+sDNA. **(B)** Results of DLS measurements on b/HLH/Z+sDNA, b/HLH+sDNA. Note the different scale of the abscissa between the first two and the other graphs. Also see Table 3. **(C)** Equilibrium analytical ultracentrifugation of the USF b/HLH/Z' construct. The data shown were collected at 280 nm with a rotor speed of 20,000 rpm, a temperature of 20 °C, in a buffer composed of 150 mM KCl, 20 mM Hepes-KOH pH 7.5, and 1 mM $MgCl_2$. The fit corresponds to a three component model with a monomer-dimer association constant of 0.8 nM and a dimer-tetramer association constant of 0.98 μM. The monomer molecular weight used in the fit was 13,500 a.m.u.


A
USF+sDNA
A/b/HLH+sDNA
Normalized Frequency (%)
20
15
10
5
0
0
500
1000
1500
2000
2500
3000
3500
Apparent Molecular Weight (kDa)
B
b/HLH+sDNA
b/HLH/Z+sDNA
Normalized Frequency (%)
60
50
40
30
20
10
0
20
25
30
35
40
45
50
Apparent Molecular Weight (kDa)

C

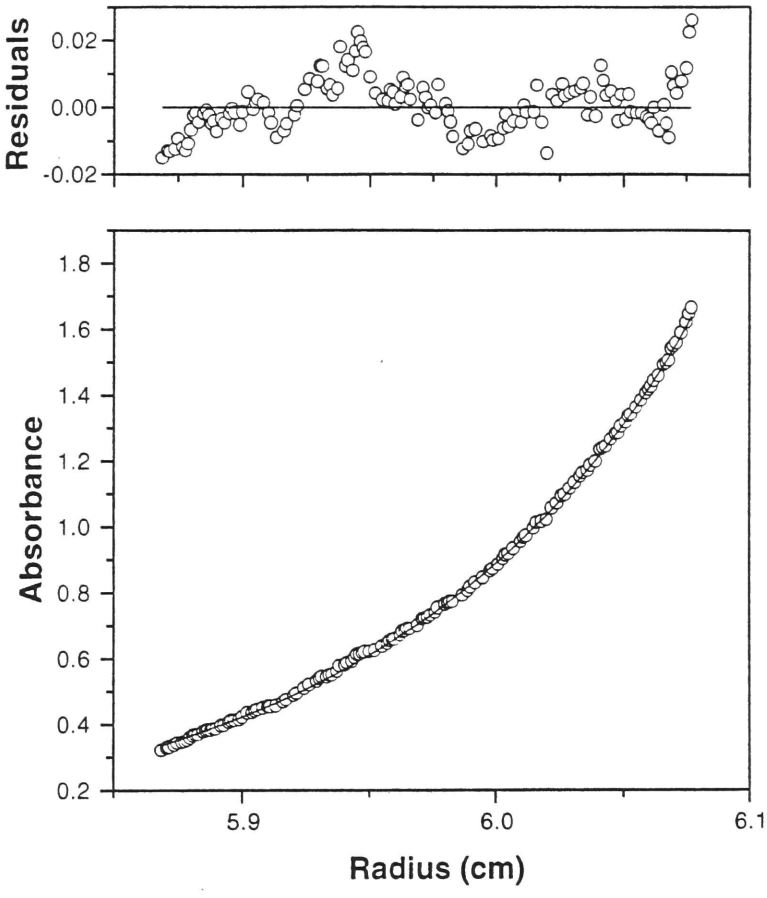

Residuals
0.02
0.00
-0.02
Absorbance
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
5.9
6.0
6.1
Radius (cm)

Table 3. Summary of dynamic light scattering results from monodisperse samples

Sample	$D_T/10^{-13}m^2s^{-1}$	R_H/nm	estimated MW/kDa	oligomerization state	calculated MW/kDa*
b/HLH/Z	737.5 (21.3)	3.35 (0.12)	55.8 (4.16)	$(b/HLH/Z)_4$	54.0
b/HLH	1055.7 (29.3)	2.18 (0.07)	19.6 (1.59)	$(b/HLH)_2$	15.2
b/HLH/Z+sDNA	614.9 (11.0)	3.74 (0.05)	72.8 (1.54)	$(b/HLH/Z)_4(DNA)_2$	73.4
b/HLH+sDNA	980.0 (15.4)	2.40 (0.04)	24.8 (1.04)	$(b/HLH)_2DNA$	24.9
Max1-113+sDNA	645.7 (26.8)	3.51 (0.20)	62.5 (9.41)	$(Max1\text{-}113)_4(DNA)_2$	67.4
Max22-113+sDNA	683.2 (4.88)	3.35 (0.05)	55.5 (0.85)	$(Max22\text{-}113)_4(DNA)_2$	62.6
c/sSREBP+SRE20	620.9 (11.0)	3.55 (0.05)	65.2 (1.87)	$(c/sSREBP)_4(DNA)_2$	64.7

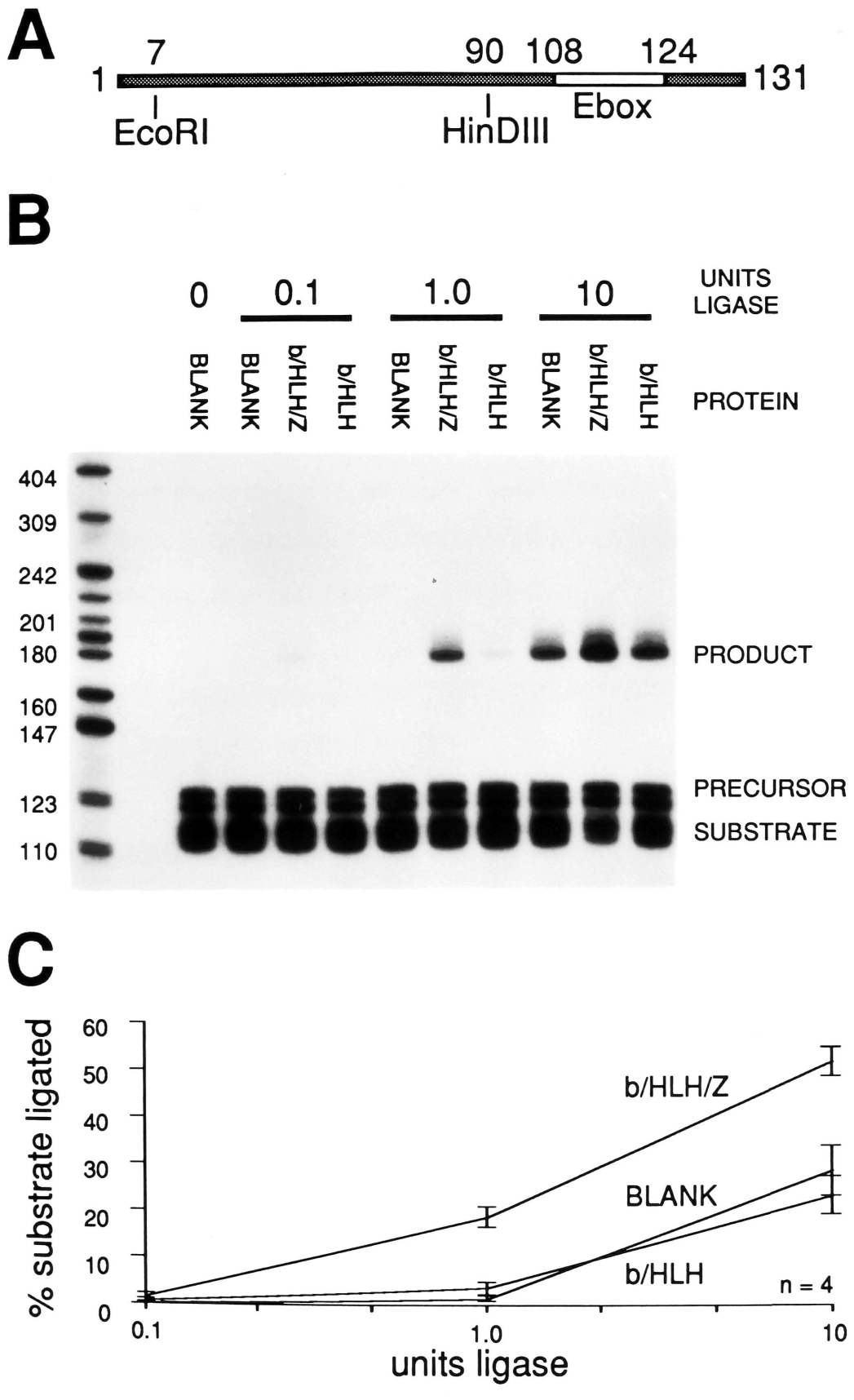
Averages (and standard deviations) are given for the translational diffusion coefficient (D_T), the equivalent hydrodynamic radius of gyration (R_H), and the estimated apparent molecular weight. Also shown are the deduced oligomerization state, and the calculated molecular weight (see Section 2.1 for the covalent molecular weights of the proteins).

* The molecular weight of sDNA is 9.7 kDa, that of SRE20 is 12.2 kDa.

The sensitivity of light scattering techniques to larger particles (reviewed in Schmitz, 1990) results in the highest order oligomer of a sample dominating the scattering signal. Therefore, a significant amount of lower oligomers could be escaping detection. In addition, if the samples are unfolded, that is, behaving as non-draining coils, their diffusion coefficients would correspond to a spuriously large mass. The compactness of the protein-DNA complexes evidenced both by circular dichroism spectroscopy and X-ray crystallography (see below) implies the mass estimates for the complexes to be valid. It should be borne in mind that for ellipsoids of revolution, even a 5:1 axial ratio results in a change in frictional coefficient of only 22 or 25% (for oblate and prolate ellipsoids, respectively) relative to a sphere of equivalent volume (Scheraga and Mandelkern, 1953). The CD spectroscopic evidence that b/HLH/Z and b/HLH proteins are partially unfolded in the absence of their target DNA prompted me to obtain independent evidence of their tetramerization when uncomplexed. Results of equilibrium analytical ultracentrifugation of b/HLH/Z in the absence of DNA are shown in Fig. 7B. The data are best fit by a three component model in which b/HLH/Z is present as a monomer, a dimer, and a tetramer with dissociation constants of $K_{\text{dimerization}} < 10^{-9}$ M and $K_{\text{tetramerization}} \sim 1$ μM.

I designed an experiment to test whether or not b/HLH/Z could function as a bivalent homotetramer at physiologic intranuclear concentration. USF is very abundant in eukaryotic nuclei. In the HeLa nucleus the USF concentration can be estimated to be 0.5 μM, assuming homogeneous protein distribution throughout a spherical nucleus of radius 2.5 μm with 20,000 molecules/cell as shown by Sawadogo *et al.*, (1988). This estimate represents a lower bound because the distribution of USF within the nucleus may not be homogeneous. To provide independent confirmation of simultaneous binding to two spatially separate DNA sites by the tetrameric leucine-zipper containing b/HLH/Z protein at a concentration of 0.5 μM, I studied the effects of exogenously added proteins on the rate of bimolecular ligation of a long duplex DNA with a USF binding site at one end and at the

Figure 8. Bimolecular ligation experiment. (**A**) Schematic depiction of the DNA employed. (**B**) Representative example of the experiment. Size marker lengths (in bases) are indicated on the left. The migration positions of ligation product, PCR product precursor and *Eco*RI-cleaved PCR product substrate are indicated on the right. All ligation reactions contained 1 μM DNA, 50 μg/ml BSA and 2 μM (calculated with the monomer molecular mass) b/HLH/Z of b/HLH where indicated. The ligation product (excised from polyacrylamide gels) was characterized by cleavage with *Eco*RI and *Hin*dIII enzymes, and demonstrated to consist exclusively of dimers of the substrate ligated through the *Eco*RI site (not shown). The ligation product migrates at an anomalously fast rate on 8 M urea gels, presumably because its being a strict palindrome favors hairpin formation. (**C**) Average of four independent experiments as shown in B. Error bars represent standard errors of the mean.

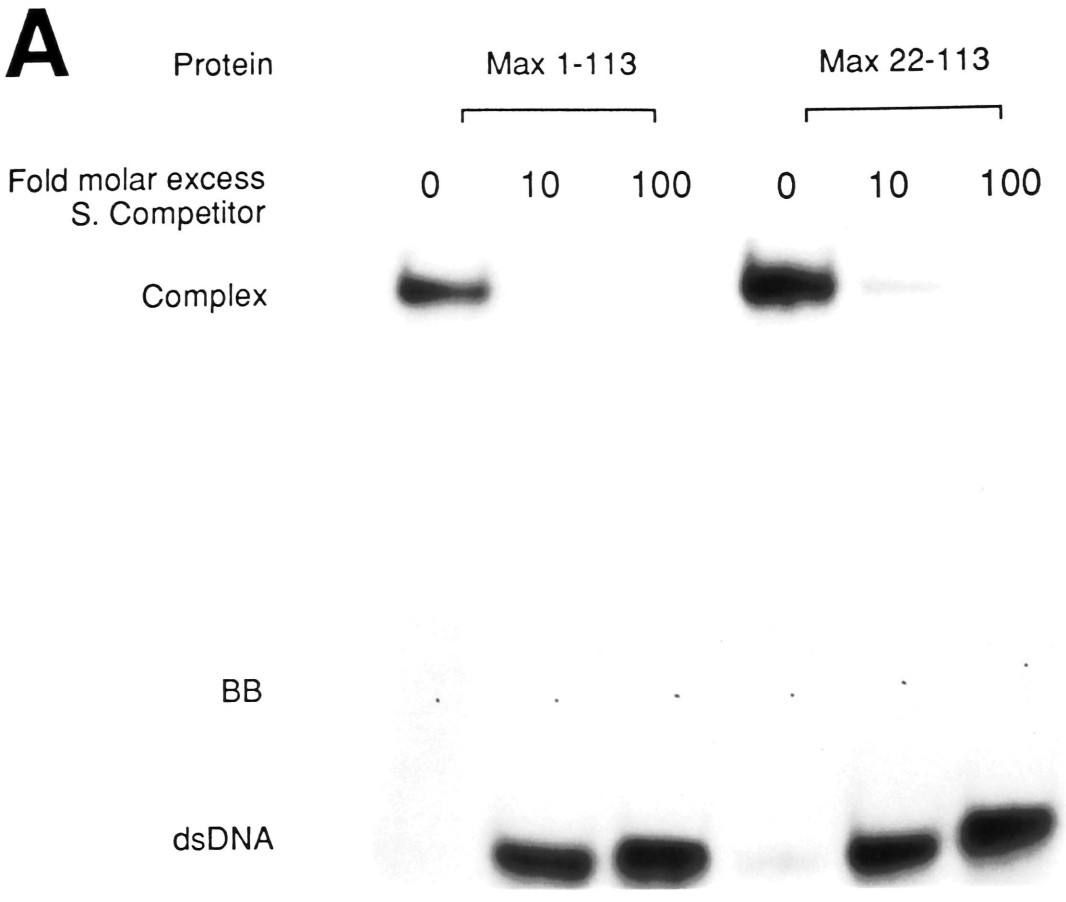

A
7
90 108 124
1
131
EcoRI
HinDIII
Ebox
B
0
0.1
1.0
10
UNITS LIGASE
BLANK
BLANK
b/HLH/Z
b/HLH
BLANK
b/HLH/Z
b/HLH
BLANK
b/HLH/Z
b/HLH
PROTEIN
404
309
242
201
180
160
147
123
110
PRODUCT
PRECURSOR
SUBSTRATE
C
% substrate ligated
60
50
40
30
20
10
0
b/HLH/Z
BLANK
b/HLH
n = 4
0.1
1.0
10
units ligase

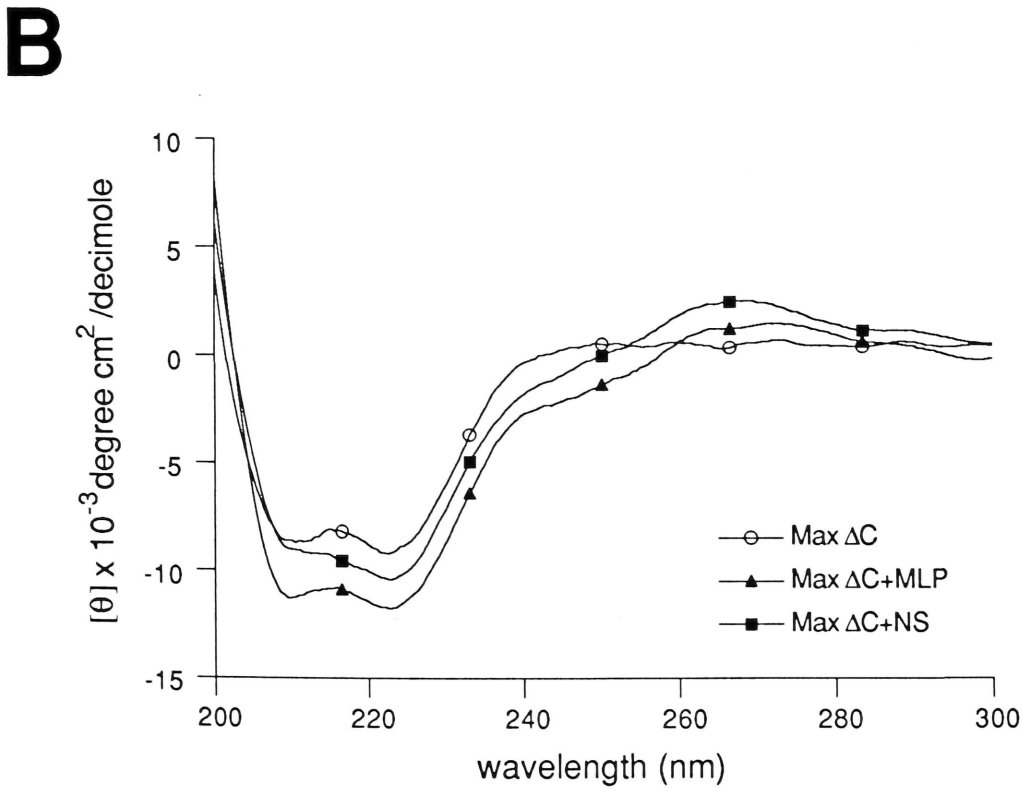
other a cohesive end generated by *Eco*RI cleavage (Fig. 8A). After purification, the ligation substrate was incubated either with a blank buffer containing a high concentration of bovine serum albumin (BSA), or with the same buffer containing either b/HLH or b/HLH/Z at a protein-DNA complex concentration of 0.5 μM. DNA ligase was then added, the reaction mixtures incubated for a fixed time and the ligation product and the substrate resolved by denaturing polyacrylamide gel electrophoresis. Results of ligation reactions performed with three different concentrations of DNA ligase are shown in Fig. 7B. Average quantities derived from four independent, replicate experiments are plotted in Fig. 7C. Presence of the tetrameric b/HLH/Z protein in the ligation reaction greatly enhances the rate of bimolecular ligation of the substrate over the rate observed in the control experiment where only BSA is present. In contrast, and as expected, the rate of bimolecular ligation in the presence of the dimeric b/HLH protein does not differ, to within experimental precision, from that observed with BSA alone.

3.2 Crystal Structure Determination of the Max b/HLH/Z Dimer in Complex with DNA

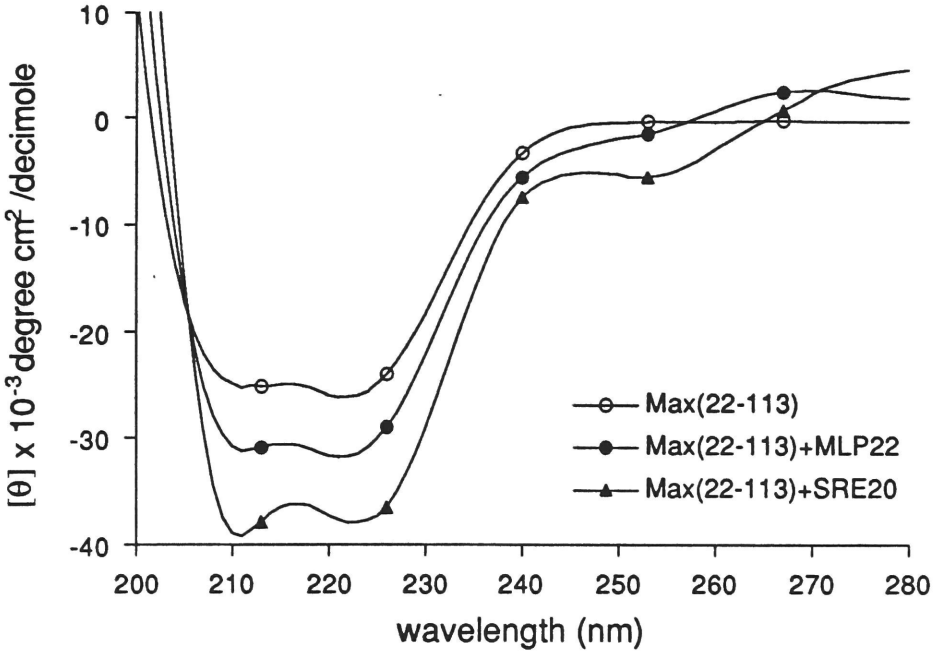
Biochemical Characterization of Max Proteins In parallel with biochemical characterization and crystallization efforts with various USF constructs, work was carried out on a second HLH transcription factor: Max. Three different Max constructs were expressed and purified to homogeneity (Chapter 2), and their biochemical properties investigated by means similar to those employed for USF. All three constructs contain the whole consensus b/HLH/Z element, and constitute functional, high-affinity DNA-binding domains, as judged by EMSA (Fig. 9A). The protein preparations also had unit activity in EMSA experiments to within experimental precision. CD spectroscopic investigation of Max 3-113 (with the hexahistidine tag removed) and Max 22-113 (Figs. 9B and 9C) show

Figure 9. EMSA and CD characterization of Max constructs. (**A**) Autoradiogram of an EMSA experiment with Max1-113 and Max22-113. Protein (on a monomer basis) and sDNA were present at a concentration of 1 μM in all reactions. Cold competitor was added in 10- and 100-fold molar excess where indicated. The positions of complex, free double-stranded DNA (dsDNA) and tracking dye (BB, bromophenol blue) are indicated. There was no detectable radioactivity retained in the loading well (not shown). (**B**) Normalized CD spectra of tag-free Max 3-113 (MaxΔC), Max3-113 with 1/2 molar equivalents of sDNA (MaxΔC+MLP), and Max3-113 with a 1/2 molar equivalents of nsDNA (MaxΔC+NS). These spectra were obtained with the samples in the phosphate/fluoride buffer system described in Section 2.4. The protein monomer concentration was approximately 6 μM for each experiment. (**C**) Normalized CD spectra of Max22-113 [Max(22-113)], this protein mixed with 1/2 molar equivalents of the major late promoter 22-mer (Fig. 4C) in complex with which its structure was determined [(Max(22-113)+MLP22], and the same protein mixed with 1/2 molar equivalents of the sterol response element 20-mer (Fig. 4F). For all three spectra, the protein concentration was approximately 14 μM, on a monomer basis. (**D**) Normalized CD spectra of the two DNAs employed in (C). The same normalization factor employed for the protein-DNA mixtures in the previous section was used for scaling these spectra.


A
Protein
Max 1-113
Max 22-113
Fold molar excess
S. Competitor
0
10
100
0
10
100
Complex
BB
dsDNA


B
10
5
0
-5
-10
-15
200
220
240
260
280
300
[θ] x 10-3degree cm2 /decimole
wavelength (nm)
Max ΔC
Max ΔC+MLP
Max ΔC+NS

C


10
0
-10
-20
-30
-40
200
210
220
230
240
250
260
270
280
wavelength (nm)
[θ] x 10-3 degree cm2 /decimole
Max(22-113)
Max(22-113)+MLP22
Max(22-113)+SRE20

D

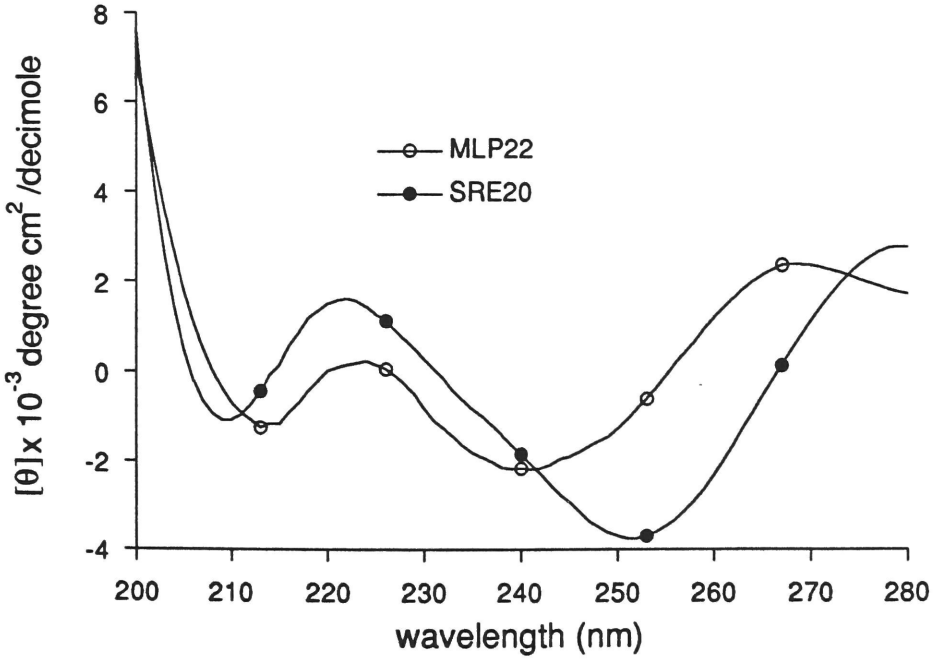
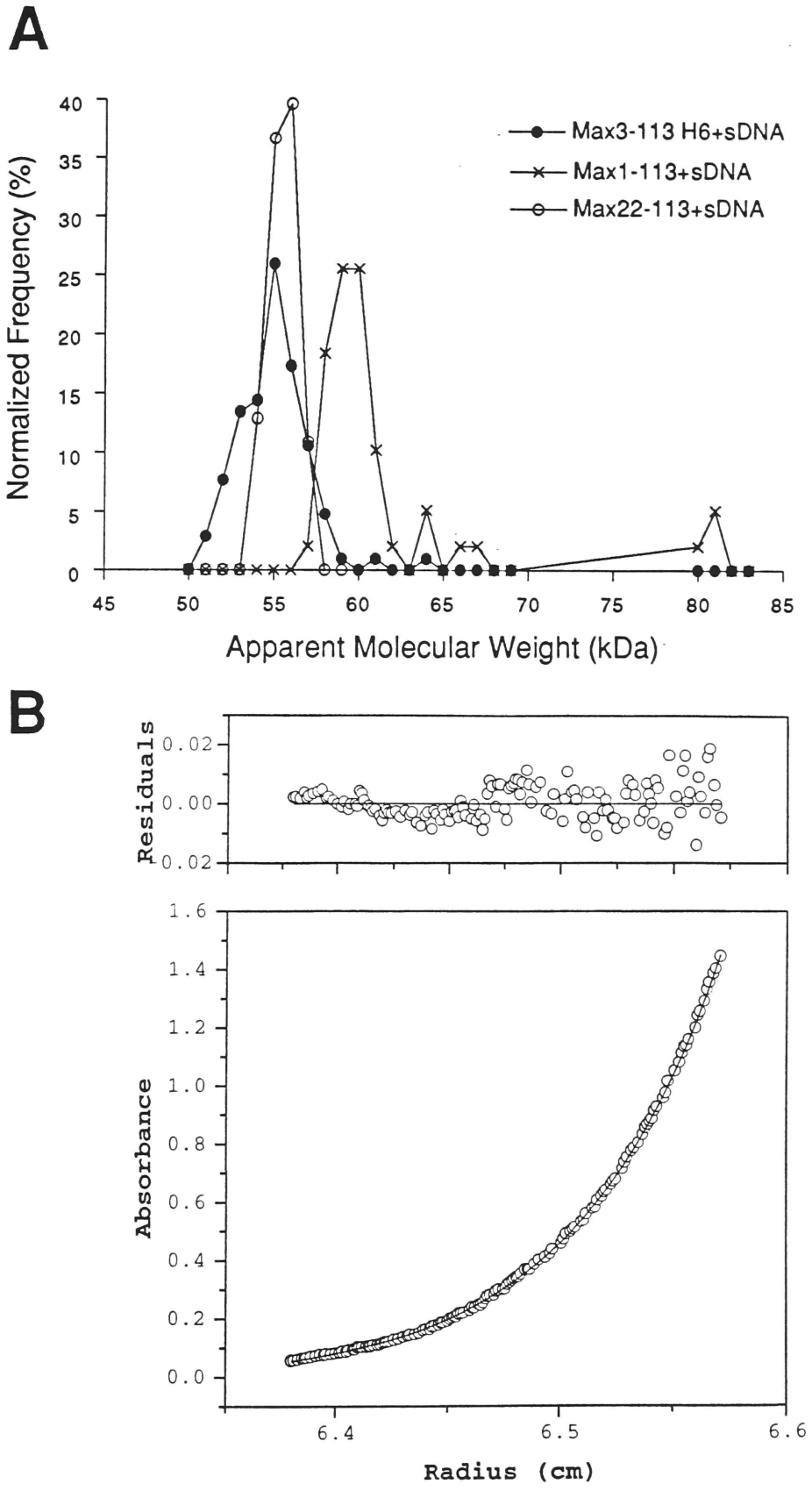

8
6
4
2
0
-2
-4
200
210
220
230
240
250
260
270
280
wavelength (nm)
[θ]x 10-3 degree cm2 /decimole
MLP22
SRE20

Figure 10. Hydrodynamic characterization of Max constructs. **(A)** Results of dynamic light scattering measurements on Max 3-113 H6+sDNA, Max1-113+sDNA, Max22-113+sDNA. **(B)** Equilibrium analytical ultracentrifugation of the Max 22-113 construct. The data shown were collected at 280 nm with a rotor speed of 30,000 rpm, a temperature of 20 °C, in a buffer composed of 100 mM KCl and 5 mM Hepes-KOH pH 7.5. The fit corresponds to a three component model with a monomer-dimer association constant of 7.3 nM and a dimer-tetramer association constant of 200 nM. The monomer molecular weight used in the fit was 10,500 a.m.u.


A
Normalized Frequency (%)
40
35
30
25
20
15
10
5
0
45
50
55
60
65
70
75
80
85
Apparent Molecular Weight (kDa)
Max3-113 H6+sDNA
Max1-113+sDNA
Max22-113+sDNA
B
Residuals
0.02
0.00
0.02
Absorbance
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
6.4
6.5
6.6
Radius (cm)

that these proteins also undergo a folding transition similar to that observed with USF b/HLH/Z.

Investigation of the hydrodynamic properties of Max 22-113 by equilibrium analytical ultracentrifugation produced results similar to those obtained with the homologous USF b/HLH/Z construct: the data are best fit by a three component model in which Max 22-113 is present as a monomer, a dimer, and a tetramer with dissociation constants of $K_{dimerization}$ = 7.3 nM and $K_{tetramerization}$ = 200 nM (Fig. 10B). DLS measurements were performed on these proteins free (not shown) and in complex with DNA. As with USF, the larger species dominate the scattering signal, and the apparent molecular masses closely approximate that expected of bivalent tetramers. Interestingly, the conformational homogeneity of the protein-DNA complexes in solution is strikingly different, with Max 22-113 producing a monodisperse solution while Max 1-113 and 3-113-H_6 aggregate considerably (Fig. 10A; Table 3). Cocrystallization experiments were carried out with all three proteins, but rapid success in growing large crystals with Max 22-113 led me to focus my efforts on this protein.

Crystal Properties and Data Collection Max 22-113 was successfully cocrystallized with a 22 base-pair oligonucleotide whose sequence was derived from the adenovirus major late promoter sequence (Section 2.8; Fig. 4C.) Under optimized conditions, these crystals grew rapidly to attain maximum dimensions of up to several mm in 8-12 hours. The morphology of the crystals is that of hexagonal prisms which taper to a sharp end (Fig. 11A.) The crystals are somewhat fragile mechanically, requiring considerable care to be mounted in capillaries for X-ray diffraction experiments. This was invariably done directly from the crystallization drops (typically 10 µl) as no artificial mother liquor was devised; reservoir solution was used for the solvent plugs. Fresh crystals diffracted isotropically with a monochromated laboratory X-ray source to about 3.2 Å resolution; however, the

Figure 11. Sample HLH-DNA cocrystals and diffraction patterns. (**A**) Example of a cocrystal of Max22-113 complexed with the MLP 22-mer DNA grown as described in Section 2.8. The crystal shown had a length of approximately 2 mm. (**B**) Diffraction pattern of a cocrystal like (A). The image was recorded on a Fuji imaging plate (IP) at beam-line F1 of CHESS. The rotation axis is horizontal, and the long axis of the hexagonal prism-shaped crystal was approximately parallel to it. The IP to crystal distance was 247 mm; the X-rays employed had a wavelength of 0.908 Å; the oscillation range was 2.5°; the total exposure time was 20 seconds. The right-hand side edge of the figure closest to the beam-stop shadow corresponds to a resolution of 3.0 Å. Note the meridional reflections at *ca.* 3.4 Å stemming from fiber-like diffraction of the partially disordered DNA in the cocrystal. (**C**) Example of a cocrystal of USF b/HLH complexed with the MLP 21-mer DNA grown as described in Section 2.8. The crystal shown had approximate dimensions 0.3 x 0.3 x 0.2 mm^3. (**D**) Diffraction pattern of a cocrystal like (C). The image was recorded on a Fuji imaging plate (IP) at beam-line F1 of CHESS. The rotation axis is horizontal, and the longest unit cell edge was approximately perpendicular to it. The IP to crystal distance was 247 mm; the X-rays employed had a wavelength of 0.908 Å; the oscillation range was 1.0°; the total exposure time was 10 seconds. The right-hand side edge of the figure closest to the beam-stop shadow corresponds to a resolution of 2.9 Å. (**E**) Example diffraction pattern of a crystal like (C) obtained with the X-rays approximately parallel to the longest unit-cell edge. Note the two sets of meridional reflections. This image was obtained with a Fuji IP at beamline X-25 of the NSLS.

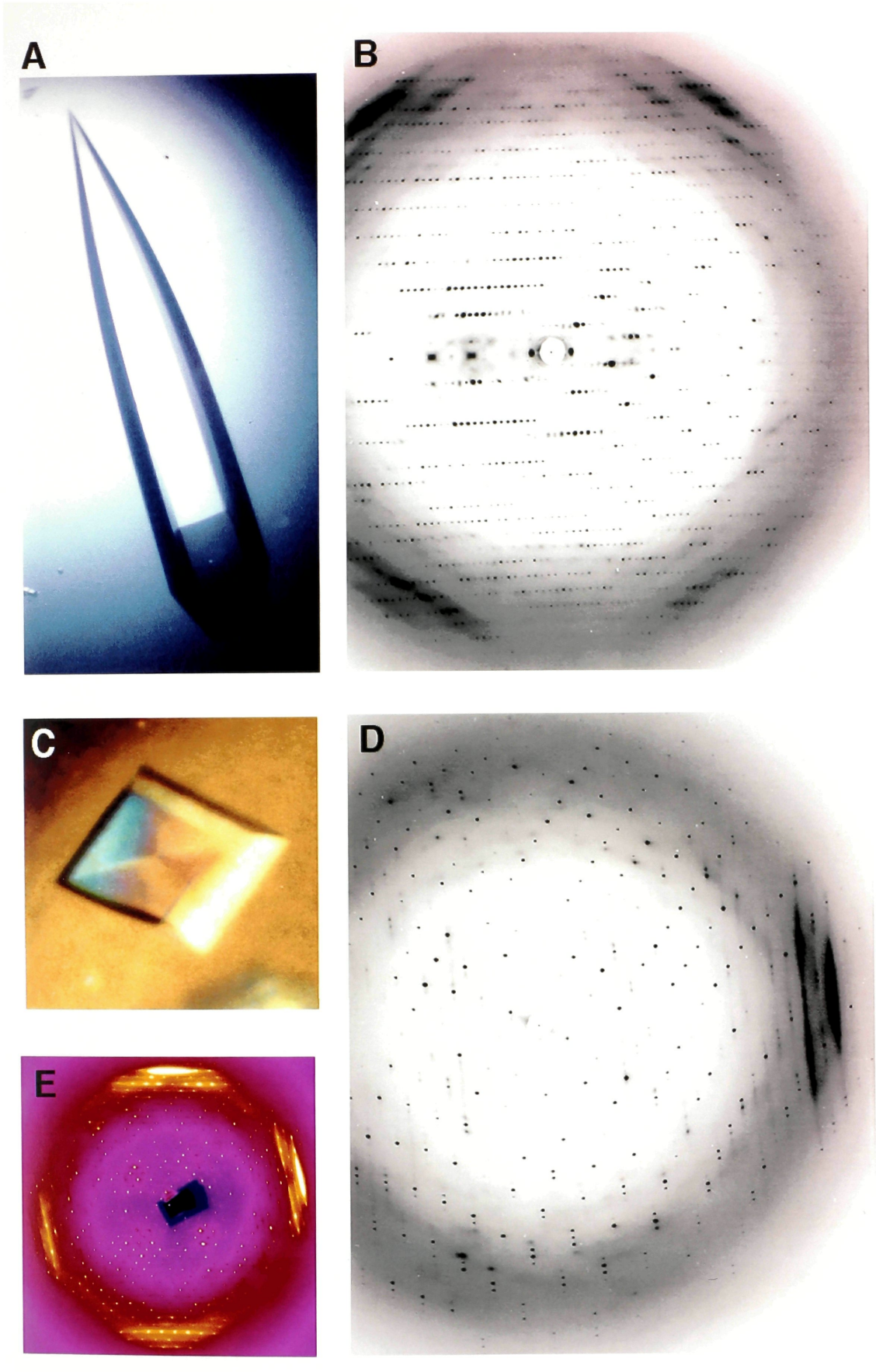

A
B
C
D
E

Table 4. Data statistics

	Native 1	Native 2	IdU4R	IdU8L	IdU8l+10L	IdU7'L	IdU10'R
Wavelength (Å)	1.5418	0.908	1.5418	1.5418	1.5418	1.5418	1.5418
Resolution range (Å)	40-3.3	50-2.9	40-3.3	40-3.3	40-3.3	40-3.3	40-3.3
Highest resolution shell (Å)	3.4-3.3	3.0-2.9	3.4-3.3	3.4-3.3	3.4-3.3	3.4-3.3	3.4-3.3
Reflections, observed	32646	21448	34661	23645	32012	30972	30512
Reflections, unique	3652	4507	3592	3554	3651	3683	3657
Data coverage (%)	96.1	82.0	94.5	93.5	96.1	96.9	96.2
Data coverage, last shell (%)	93.8	60.5	92.4	93.5	94.1	94.7	93.5
R_{sym} ($\|F\|>0$) (%)*	8.5	6.5	10.1	11.8	10.9	8.7	11.6
R_{sym} ($\|F\|>0$), last shell (%)	22.0	27.2	21.8	23.8	24.3	21.3	25.5
$\langle I/\sigma(I)\rangle$	14.7	14.4	12.4	9.65	12.5	13.5	12.7
$\langle I/\sigma(I)\rangle$, last shell	6.24	2.88	5.79	4.7	5.98	6.54	5.90
Mean fractional isomorphous difference 15-3.3Å *vs.* Native 1 (%)			16.5	13.0	10.3	9.1	10.3
Mean fractional isomorphous difference 15-3.3Å *vs.* Native 1, last shell (%)			18.3	16.1	11.7	12.6	14.2

* $R_{sym} = \Sigma|I-\langle I\rangle|/\Sigma I$, where I is the observed intensity and $\langle I\rangle$ is the average intensity obtained from multiple observations of symmetry related reflections.

crystals were found to be very radiation sensitive, with the maximum resolution decaying to less than 5 Å after 4 hours of continuous exposure to X-rays. This problem was overcome partially by cooling the crystals to -15 °C, a temperature just above their freezing point. At this temperature, the crystals withstood irradiation for approximately 20 hours without evidence of significant decay, as judged by visual inspection of oscillation photographs and by analysis of the reduced data; all subsequent diffraction experiments with these crystals were carried out at this temperature.

Collection of oscillation photographs demonstrated the presence of a long crystallographic unit cell axis coincident with the long morphological dimension of the crystals (Fig. 11B.) This made it expedient to mount these crystals in capillaries with this long axis aligned with the camera spindle axis. Rotation about this axis brought zones into the diffracting position every 60°, suggesting that the crystals were in either a hexagonal or a trigonal point group. The unit cell was determined both manually, by indexing oscillation photographs employing the interactive graphics version of DENZO, as well as with the auto-indexing and cell reduction programs of the Molecular Structure Corporation R-AXIS data reduction software. The refined unit cell had dimensions of a = 72.2 Å and c = 146.4 Å, (with $\alpha = \beta = 90°$ and $\gamma = 120°$). Oscillation photographs were collected over a 60° range, with an exposure time of 20 minutes/degree, a crystal to detector distance of 180 mm, and an oscillation range of 2° per photograph. Images were reduced employing DENZO, and merged and scaled with SCALEPACK. Initially the data were merged in space groups *P*3, *P*32, *P*6, and *P*622. The resulting number of rejected observations (out of a total of *ca.* 30000 observations) and final merging *R*-factors were (167, 7.0%), (167, 7.0%), (165, 7.5%), and (181, 8.1%), respectively for the four space groups. The similarity in these numbers implied that the crystals were likely to have 622 point-group symmetry.

The hexagonal unit cell has a volume of 7.64 x 10^5 $Å^3$. Assuming 622 point-group symmetry, a full complex (the molecular weights of the DNA strands are 6807 and 6616 Da, of a protein monomer 10826 Da) per asymmetric unit gives a Matthews number (Matthews, 1968) of 1.81 $Å^3$/Da; half a complex per asymmetric unit 3.62 $Å^3$/Da. Given the fragility of the crystals it was deemed more likely that the latter was the case: the asymmetric unit would then include one Max 22-113 polypeptide and only 11 of 22bp of the quasisymmetric DNA. Because the accumulated biochemical evidence on HLH protein-DNA interaction implied that these proteins bound symmetrically to the dyad symmetric E-box (CACGTG), it was decided to proceed to attempt to solve the structure of this complex, despite the two-fold averaging of the DNA. (Attempts at growing crystals under similar conditions with a fully symmetrical DNA sequence resembling the adenovirus major late promoter sequence were, unexpectedly, unsuccessful.)

A similar situation has been described by DiGabriele *et al.* (1989). These investigators found that a DNA dodecamer containing an "A-tract" crystallized in both orientations, with approximately half the molecules in the crystal pointing in one direction. Because the entire DNA molecule constituted the asymmetric unit, introduction of a single bromine in the DNA duplex resulted in two peaks in isomorphous difference Fourier syntheses. In the case of the Max-DNA complex, only half the DNA is present in each asymmetric unit; introduction of single iodine in the derivative DNAs resulted in peaks with half-occupancy, and correspondingly low mean fractional isomorphous differences and phasing powers.

Diffraction data from one native and five derivative crystals, extending to 3.3Å resolution, were collected employing a laboratory X-ray source, and reduced with DENZO and SCALEPACK. A single crystal was employed per data-set. One higher resolution native data-set was collected from one crystal at beamline F-1 of the Cornell High-Energy Synchrotron Source. This data-set was collected with X-rays of wavelength 0.908 Å, a

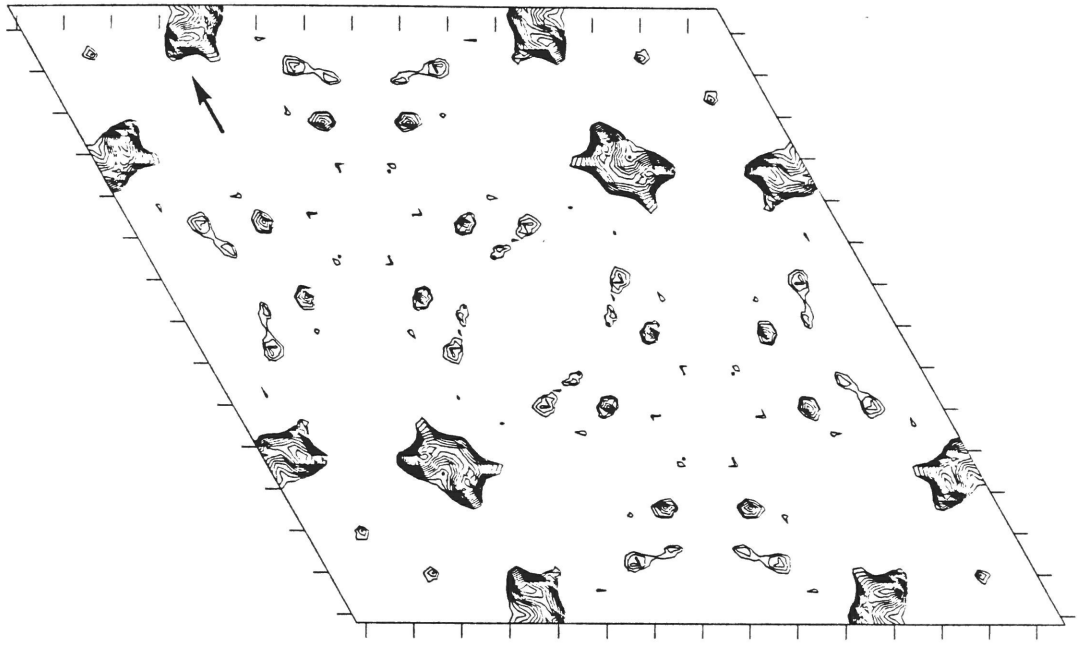


Figure 12. The w = 1/6 Harker section of the isomorphous difference Patterson synthesis for derivative IdU4R. The map is contoured at 0.25 σ intervals starting at 1.0 σ above mean peak height. The origin is at the upper left; the section shown extends slightly beyond one unit-cell. The arrow points to the peak corresponding to coordinates (y, x-y, 1/6).

Table 5. Phasing statistics

Resolution	15-3.3Å	9.8Å	6.9Å	5.6Å	4.8Å	4.3Å	3.9Å	3.6Å	3.3Å
Reflections	3599	247	337	390	448	504	518	576	579
Acentric reflections	3394	229	313	370	427	474	488	545	548
Figure of merit	0.58	0.68	0.79	0.75	0.69	0.58	0.46	0.45	0.45
IdU4R									
Phasing power*	0.93	0.75	1.25	1.02	1.00	0.84	0.87	0.93	0.91
Centric R-factor	0.75	0.70	0.63	0.71	0.53	0.90	0.91	0.87	0.83
IdU8L									
Phasing power*	0.62	0.48	0.96	1.01	0.77	0.60	0.57	0.52	0.48
Centric R-factor	0.70	0.73	0.60	0.49	0.62	0.76	0.76	0.85	0.68
IdU8l+10L									
Phasing power*	0.83	0.59	0.94	1.40	1.36	1.13	0.78	0.69	0.60
Centric R-factor	0.64	0.70	0.58	0.43	0.59	0.71	0.73	0.73	0.71
IdU7'L									
Phasing power*	0.76	0.60	1.16	1.20	1.23	0.80	0.58	0.61	0.63
Centric R-factor	0.70	0.57	0.54	0.64	0.84	0.78	0.79	0.77	0.86
IdU10'R									
Phasing power*	0.71	0.49	0.93	1.17	1.07	0.88	0.69	0.58	0.53
Centric R-factor	0.70	0.68	0.67	0.48	0.56	1.05	0.81	0.71	0.65

* Phasing power, r.m.s.($|F_H|/E$); $|F_H|$, heavy atom structure factor amplitude, and E, resudual lack of closure error.

crystal to detector distance of 247 mm, an oscillation range of 2.5° per photograph, and an exposure time of 10 seconds/degree. Six exposures were made, then the crystal was translated parallel to the spindle axis, and six more exposures were made. This gave a total of 30° of data from a single crystal in under five minutes. Fast data collection with the intense CHESS source yielded data to 2.9 Å. The refined mosaicity of this crystal (refined with SCALEPACK) was 0.4°. Statistics for all seven data-sets appear in Table 4. The values of the merging reliability index (R_{sym}) for the native data-sets are similar to what has been reported with the imaging plate detector system on protein crystals (4-8%; Sato *et al.*, 1992; Gruner, 1994); the derivative data-sets have somewhat larger, but not unreasonable, values of this statistic, presumably a result of increased radiation sensitivity due to the presence of the halogen.

Phase Determination The structure was solved at 3.3Å resolution by multiple isomorphous replacement using the five iodinated DNA derivatives. The heavy atom site for derivative IdU4R, the derivative with the largest mean fractional isomorphous difference, was located by difference Patterson analysis. Fig. 12 shows the w = 1/6 section of the isomorphous difference Patterson synthesis. Harker sections were located also at w = 1/3 and w = 1/2, implying that the crystals belong to either space-group $P6_122$ or $P6_522$. The position and occupancy of the iodine from this derivative were refined employing HEAVY against isomorphous differences in space-group $P6_122$ to give an initial figure or merit of 0.19. Refinement against both isomorphous and anomalous differences gave a figure of merit of 0.36. The remaining four derivatives were characterized by difference Fourier syntheses; refinement against both isomorphous and anomalous differences for all five derivatives gave an overall figure of merit of 0.58. Phasing statistics are given in Table 5; the refined heavy-atom parameters in Table 6. The atomic coordinates of the iodines were consistent with the halogens being bonded to the 5 positions of their respective deoxyuridines in B-form DNA.

Table 6. Refined heavy atom parameters*

Derivative	x	y	z	relative occupancy
IdUR4	0.2501	0.2139	0.0175	3.521
IdU8L	0.3254	0.1121	0.3450	1.749
IdU8L+10L	0.3229	0.1118	0.3459	1.587
	0.4552	0.1418	0.3462	1.372
IdU7'L	0.4015	0.2187	0.3116	1.792
IdU10'R	0.4892	0.2674	0.3839	1.854

* *B*-factors were fixed at 20Å^2. Positions are in fractional coordinates.

Table 7. Crystallographic refinement statistics of the (Max 22-113)$_2$-DNA complex

Resolution range (Å)	6.0-2.9
R factor (%)	23.2
Reflections ($\|F\| > 1\sigma\|F\|$)	3916
Total number of non-hydrogen atoms	1165
R.m.s. bond lengths (Å)*	0.013
R.m.s. bond angles (degrees)*	2.29
R.m.s *B*-factors bonded atoms (Å^2)*	3.5
Average *B*-factor protein main chain (Å^2)†	28.5
Average *B*-factor protein side chains (Å^2)†	29.2
Average *B*-factor DNA bases (Å^2)†	12.9
Average *B*-factor DNA backbone (Å^2)†	21.1

* R.m.s. bond lengths and r.m.s. bond angles are the respective root-mean-square deviations from ideal values; r.m.s. *B*-factor is the root-mean-square deviation between the thermal parameters of coavalently bonded atomic pairs.

† Average *B*-factors are arithmetic means.

Model Building and Refinement The 3.3Å resolution MIR map revealed clear electron density for the phosphodiester backbone, most of the bases, and a long α-helix corresponding to the basic region plus H1 (b/H1, see legend to Fig. 2). "Solvent flattening" with SQUASH improved the electron density considerably. A canonical B-form DNA with the sequence shown in Fig. 4G was positioned in the electron density map by overlaying thymine methyl carbon atoms on four iodine atomic coordinates, giving a root-mean-square deviation of 0.93Å. Thereafter, base pairs were manually repositioned using O to fit the modified electron density, and the DNA model was further improved using X-PLOR positional refinement. Phase combination with the MIR data and the refined DNA model using COMBINE revealed clear electron density for the b/H1 α-helix and most of the α-helix corresponding to H2 plus the leucine zipper (H2/Z), which were initially built as two α-helical polyalanine segments. Several rounds of model building, phase combination, and refinement allowed an unambiguous trace and sequence assignment of the polypeptide chain from Alanine 22 to Serine 107 (94% of Max22-113) with an *R*-factor of 27.6% (with all *B*-factors set at 20 $Å^2$.) During the initial rounds of refinement, the sugar puckers were dynamically restrained to have a 2'-endo pucker, and some of the DNA-base pairs were restrained to maintain Watson-Crick hydrogen bonding by employing pseudo nOe constraints in X-PLOR. These restraints were gradually removed as the refinement progressed.

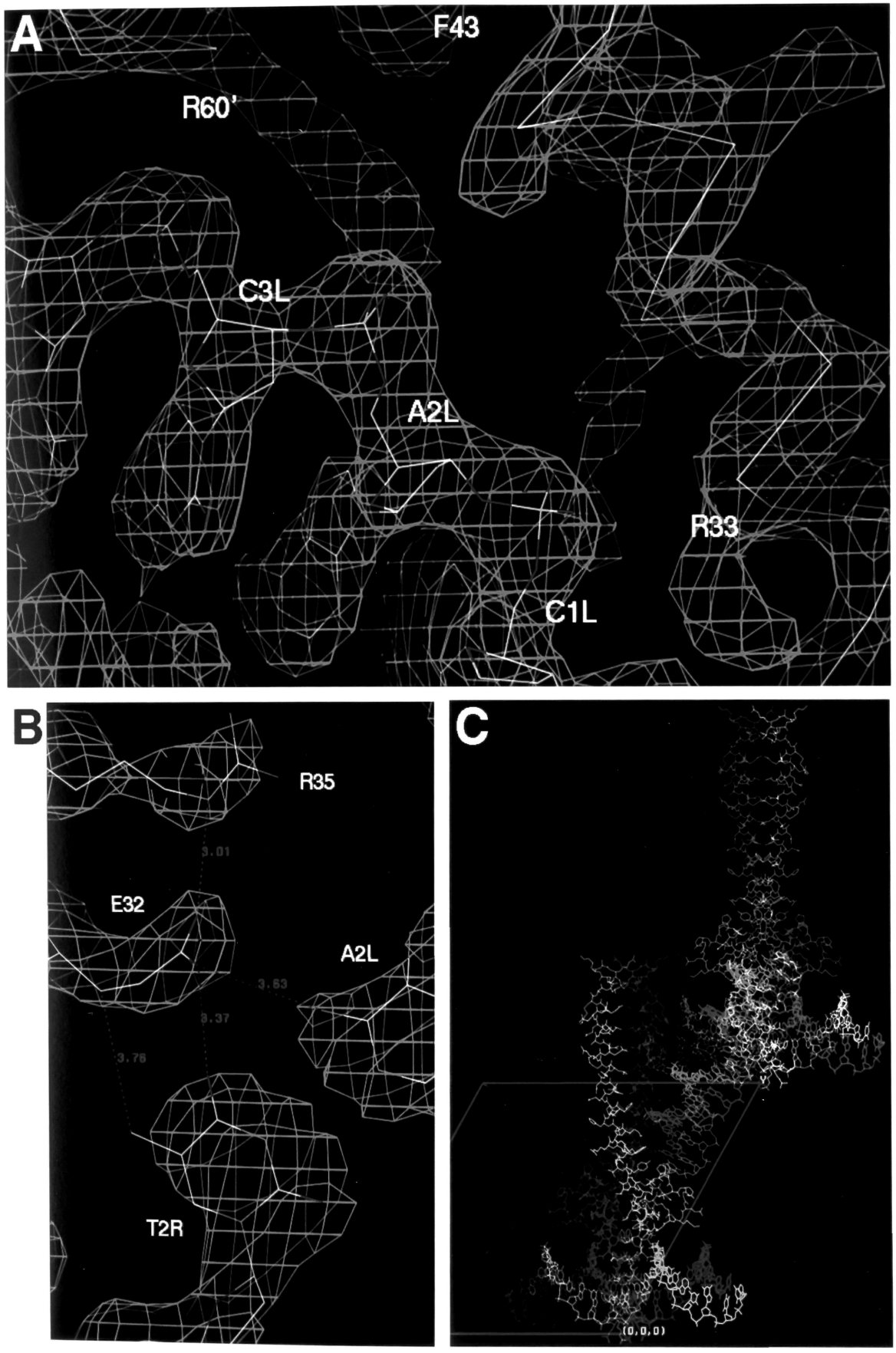
The model was then refined at 2.9Å resolution, against the "Native 2" data-set, using simulated annealing, followed by X-PLOR positional refinement. Omit difference Fourier syntheses were calculated with coefficients ($2|F_{observed}|-|F_{calculated}|$) and phases derived from the refined model (with regions of ambiguity omitted, to reduce phase bias) and several rounds of manual rebuilding and positional refinement were carried out to improve stereochemistry. Finally, tightly-restrained, individual isotropic *B*-factors were refined. The first crystallographic refinement model comprised the double-stranded DNA 11-mer

depicted in Fig. 4G and the Max22-113 amino acid sequence from Alanine 22 to Serine 107.

Subsequently, the DNA model was changed to the sequence (equivalent by symmetry to the first model) shown in Fig. 4H. This allowed incorporation of the phosphate 5' to G1R in the crystallographic refinement. This change in the model resulted in modest improvement in some parts of the electron density maps. The model was manually rebuilt accordingly and further refined using X-PLOR positional refinement followed by refinement of tightly-restrained individual isotropic *B*-factors to yield the final model. No attempt was made to place solvent molecules. Refinement statistics are given in Table 7, and portions of the final electron density are shown in Figs. 13A and 13B.

The Crystallographic Model The electron density for the polypeptide backbone for b/H1, the loop and H2 is continuous at 1σ in a ($2|F_{observed}|-|F_{calculated}|$) difference Fourier synthesis. Towards the C-terminus of the leucine zipper, where the electron density is not as well defined, there are a few density breaks along the polypeptide backbone. Omit maps were used to examine the electron density for the protein, which revealed no evidence of multiple conformations of the polypeptide backbone at this resolution limit. Some of the solvent exposed side chains showed alternate conformations, but no attempt was made to include alternate conformations in the refinement. The electron density for the central six base-pairs of the DNA is well-defined and connected for both the bases and the backbone (Fig. 13A). Beyond the recognition element, the electron density for the backbone is well connected with no evidence of significant conformational averaging. The electron density for bases in asymmetric positions is consistent with a superposition of both sequences. A Ramachandran (Ramachandran *et al.*, 1963) analysis with PROCHECK showed only 1 of 85 backbone torsion angle combinations in a disallowed region of the (ϕ,ψ) plot, five in generously allowed regions, 22 in allowed regions and 54 in most favored regions (Fig.

Figure 13. Final electron density and crystal packing of the $(Max22\text{-}113)_2$-DNA complex structure. (**A**) Electron density map calculated with $(2|F_{observed}| - |F_{calculated}|)$ coefficients and phases calculated from the final refined model. This view shows the upper part of the basic region, the upper strand of DNA adjacent to it, and the bottom residue (R60') of the H2/Z α-helix of the dyad-related protein molecule. Only the Cα positions of the protein model are shown for clarity. The contour level is 1.5 σ above mean peak height. (**B**) Electron density map calculated as (A). This view shows the A:T base pair of the E-box. The side-chains of Glutamate 32 and Arginine 35, and some of the inter-atomic distances are shown. Note that the precision of the atomic coordinates is approximately 0.35 Å (see also Fig. 15). The contour level is 1.25 σ above mean peak height. (**C**) Crystal packing. The contents of six asymmetric units are shown in different colors. Four of the asymmetric units are shown in their entirety, two have had the protein model omitted for clarity. The view is approximately parallel to the hexagonal (c = 146.4 Å) axis. The boundaries of a unit cell are shown in red lines, with the origin close to the center bottom of the figure [marked (0, 0, 0)] For reference, $a = b$ = 72.2 Å. The space group is $P6_122$; there are twelve asymmetric units per unit cell.


A
F43
R60'
C3L
A2L
R33
C1L
B
R35
3.01
E32
A2L
3.63
3.37
3.76
T2R
C
(0,0,0)

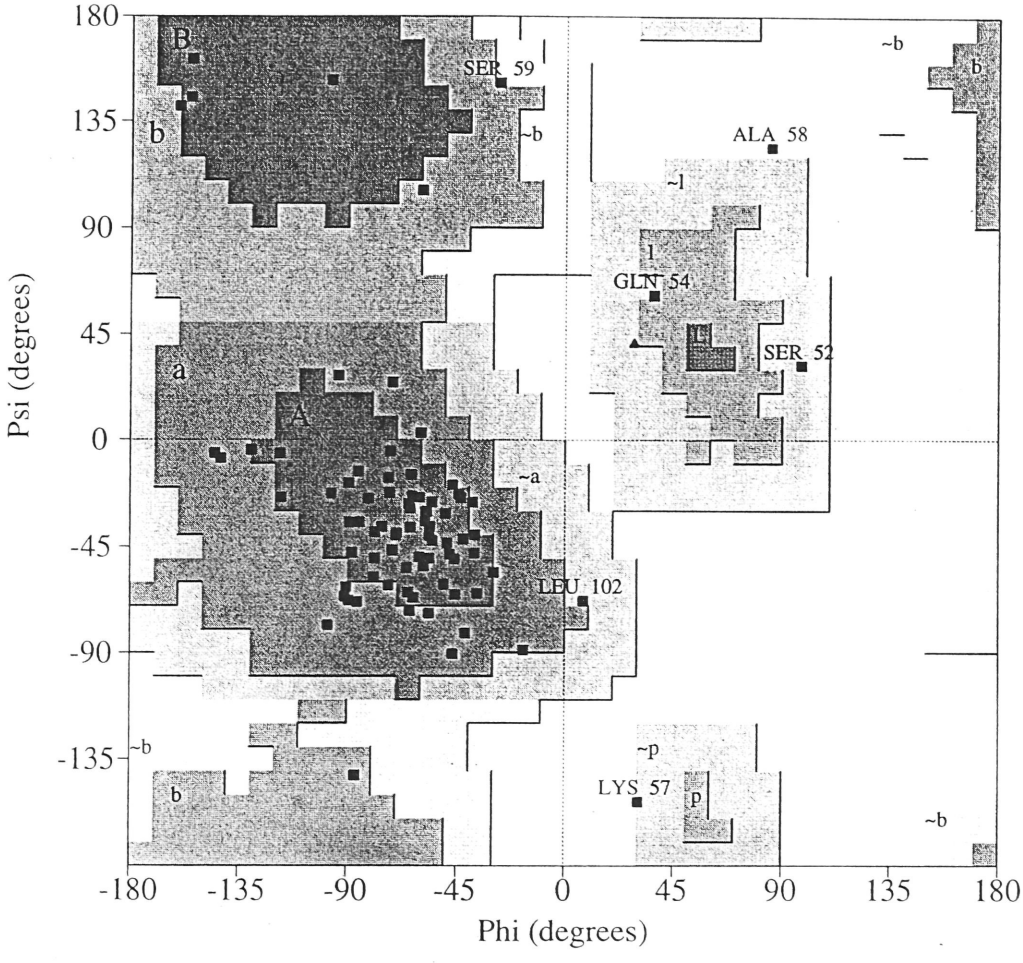


Figure 14. Ramachandran plot of the final refined (Max22-113)$_2$-DNA complex structure prepared using PROCHECK (Laskowski *et al.*, 1993). One residue, Alanine 58, is in a disallowed region of the plot, and five are in "generously allowed regions". With the exception of Leucine 102, which is at the C-terminal end of the zipper motif, all lie in the loop region.

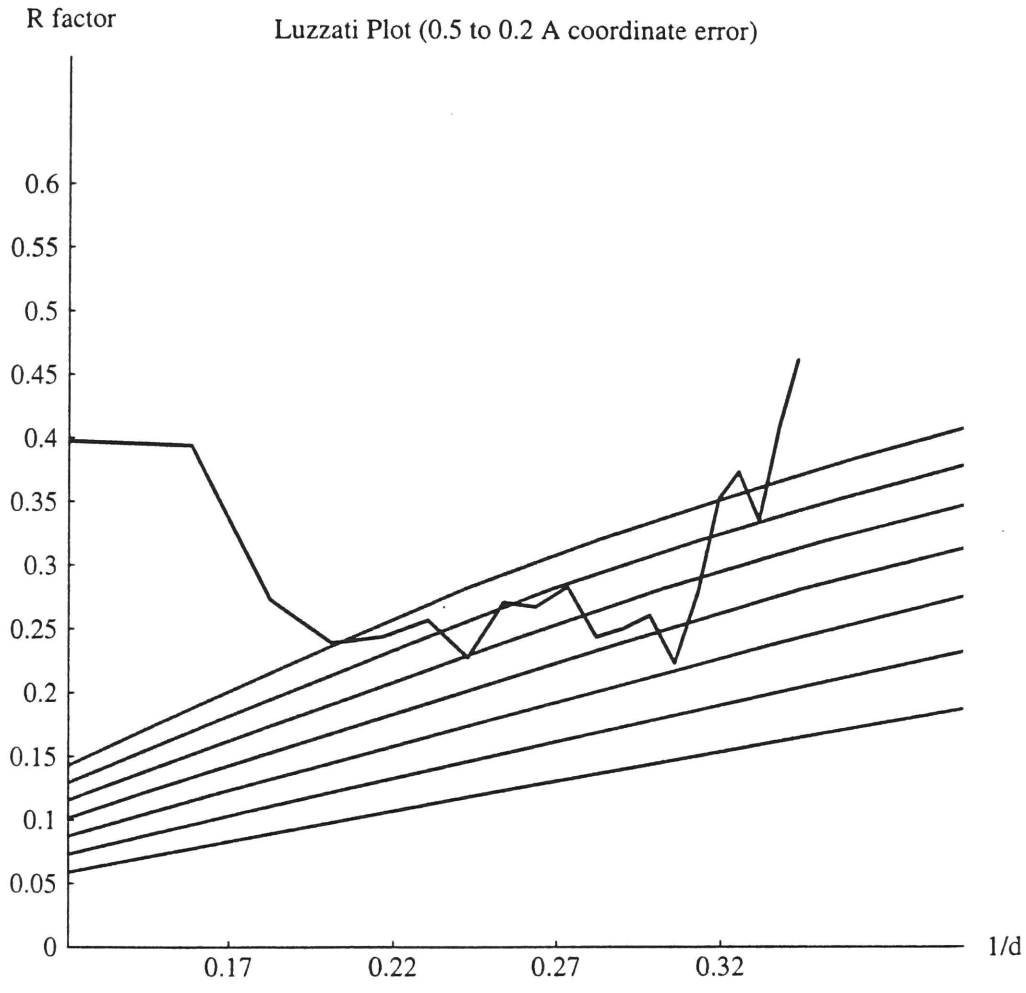
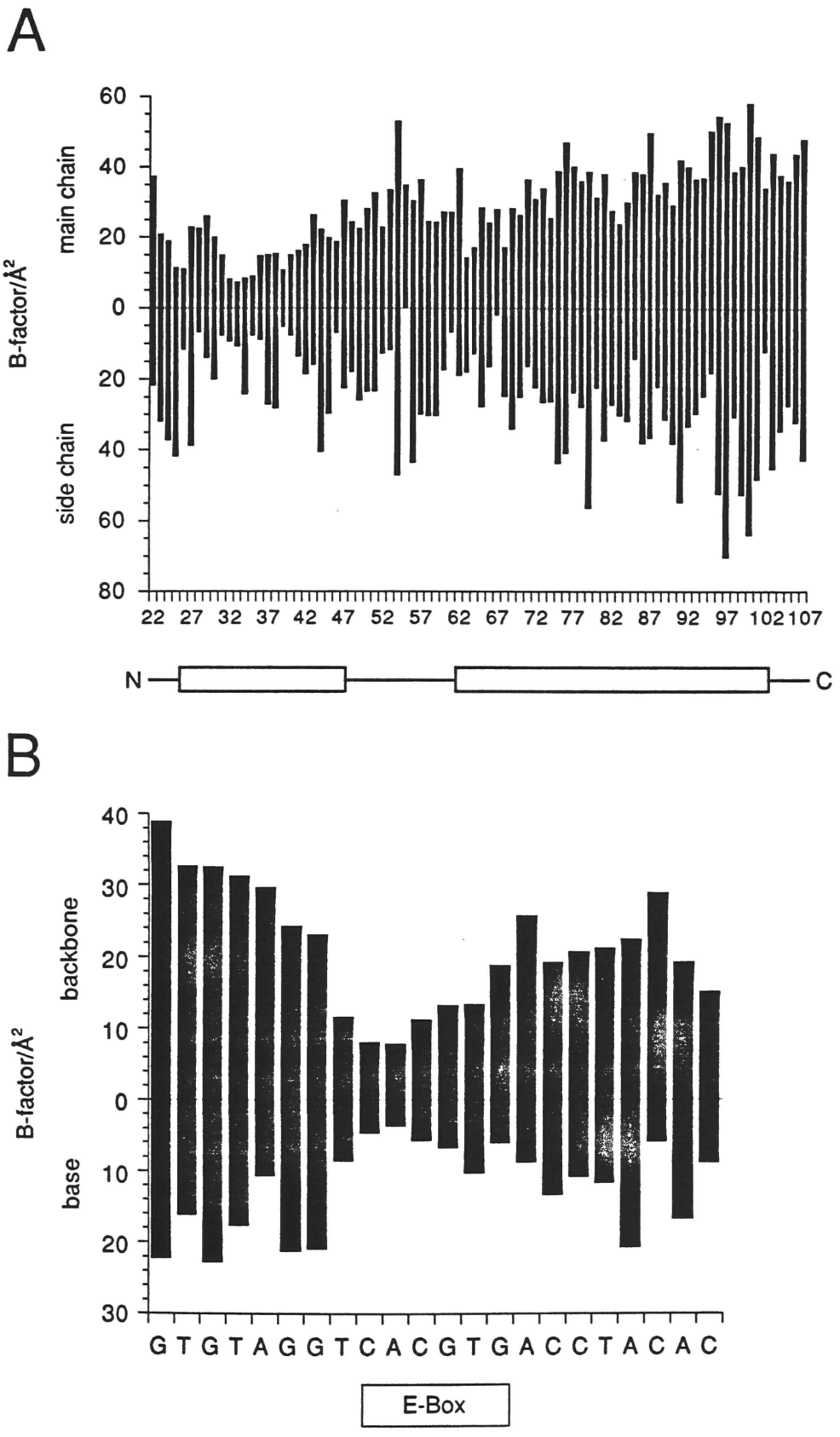


Figure 15. Luzzati plot of the final refined (Max22-113)$_2$-DNA complex structure. Diagonal curves correspond to mean coordinate errors of 0.2 to 0.5 Å in 0.05 Å intervals, from bottom to top.

Figure 16. Individual isotropic *B*-factors of the final refined (Max22-113)$_2$-DNA complex structure. The ordinates correspond to the arithmetic mean of the *B*-factors of atoms in the residue whose number is indicated on the abscissa. (**A**) *B*-factors of protein main chain and side-chain atoms in the refined model. The two helical regions of the protein are indicated as rectangles below the abscissa; regions of irregular secondary structure are symbolized with thin lines. (**B**) *B*-factors of atoms in the nucleotide bases or the phosphodiester backbone of the refined DNA model; the sequence of the DNA is the one shown in Fig. 4H. The position of the recognition element, the E-box, is shown below the abscissa.


A
B-factor/Å²
main chain
side chain
60
40
20
0
20
40
60
80
22 27 32 37 42 47 52 57 62 67 72 77 82 87 92 97 102 107
N
C
B
B-factor/Å²
backbone
base
40
30
20
10
0
10
20
30
G T G T A G G T C A C G T G A C C T A C A C
E-Box

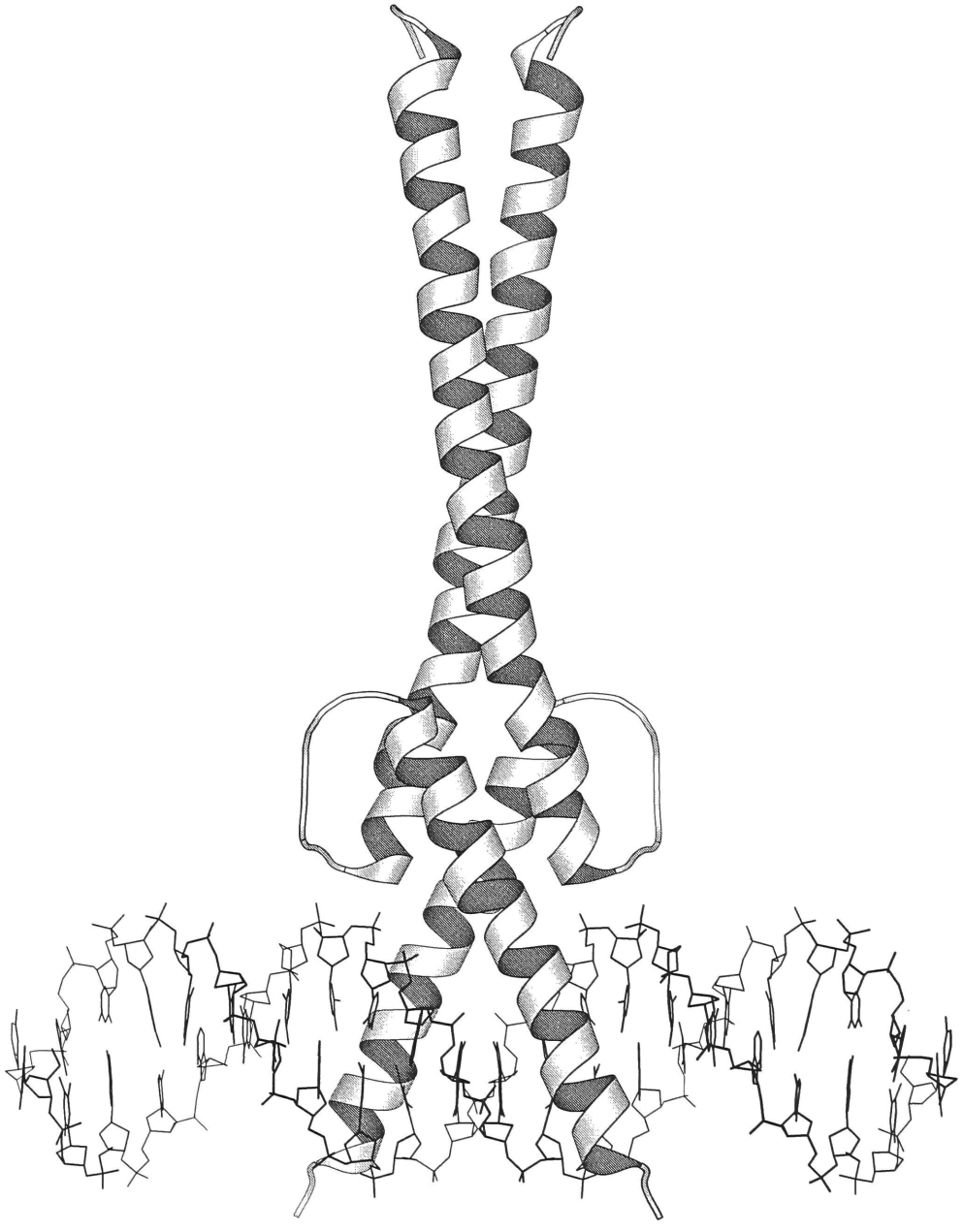
14). PROCHECK analysis of the geometry of the model indicated it to be consistent with or better than that expected at this resolution limit (not shown.) The Luzzati plot (a means of quantitatively estimating the coordinate error in the final model; Luzzati, 1952) shown in Fig. 15 indicates that the precision of the coordinates of the model is 0.3 - 0.4 Å. The *B*-factors are plotted as a function of residue number in Fig. 16. It can be seen that the protein-DNA interface (see below) is the most ordered region of the crystalline complex. The average *B*-factors of the protein, ~ 29 $Å^2$, and the DNA, ~ 17 $Å^2$, lie within the range typically observed in protein structures (Drenth, 1994; p. 94).

3.3 Structure of the Max b/HLH/Z Dimer in Complex with DNA

Topological Overview The DNA-binding domain of Max consists of two lengthy α-helices separated by a loop (Fig. 17). The N-terminal α-helix (b/H1) is continuous, and includes residues from the basic and H1 regions. The second α-helix (H2/Z), also continuous, is composed of the H2 and leucine zipper regions. Max binds DNA as a homodimer, and the two monomers fold into a parallel, left-handed, four-helix bundle. Two α-helices, the two basic regions, project from the four-helix bundle towards the DNA and enter the major groove in opposite directions. The H1 regions make up half of the four-helix bundle, packing against each other and the two H2 regions. Finally, two Z portions of the second α-helical segment form a parallel, left-handed coiled coil.

Analysis of Solvent Accessible Surface Areas Solvent-accessible surface areas (Lee and Richards, 1971) were calculated for diverse combinations of portions of the Max-DNA complex structure as a means of analyzing the architecture of the complex. Given the precision of atomic coordinates, burial of accessible surface area is probably a more meaningful criterion for analyzing the structure than are interatomic distances. Dimerization

Figure 17. Overall view of the structure of the $(Max22\text{-}113)_2$-DNA complex. Regions of the protein with α-helical structure are shown as ribbons, regions with irregular secondary structure as thin tubes. The DNA model is shown in thin lines. The N-termini of the two protomers are at the bottom of the figure; the C-termini are close to the top. The crystallographic dyad (two-fold) axis lies parallel to the plane of the paper, and bisects the complex vertically. Figure prepared with MOLSCRIPT (Kraulis, 1991).



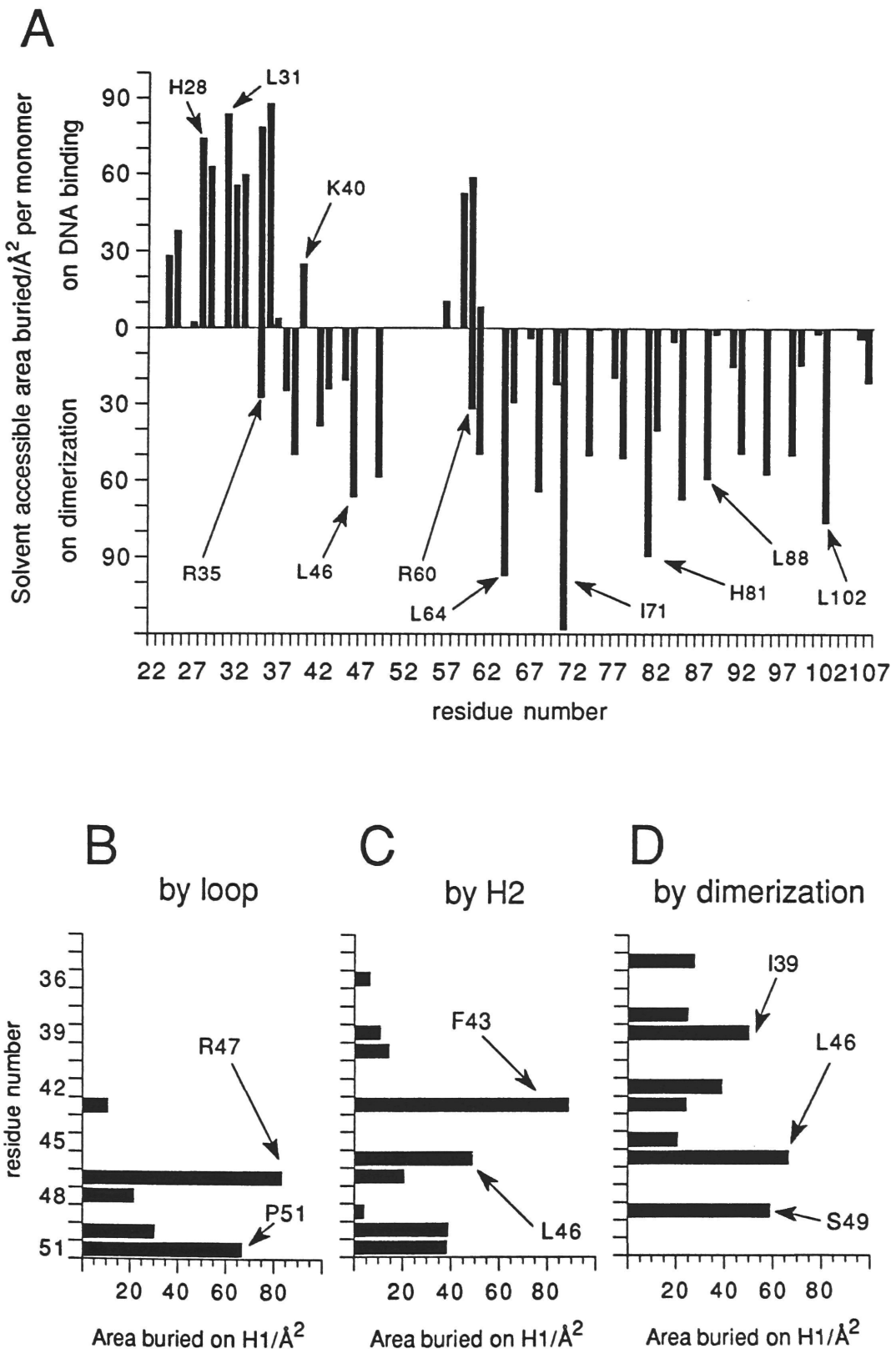
results in a reduction of solvent-accessible surface area of 1410 $Å^2$ per monomer out of a total of 7930 $Å^2$. DNA binding results in a further reduction of 720 $Å^2$ per monomer. The total solvent-accessible surface area of the free DNA is 4160 $Å^2$ per strand. For comparison, dimerization of the leucine zipper of the b/Z protein GCN4 results in the burial of 900 $Å^2$ per monomer.

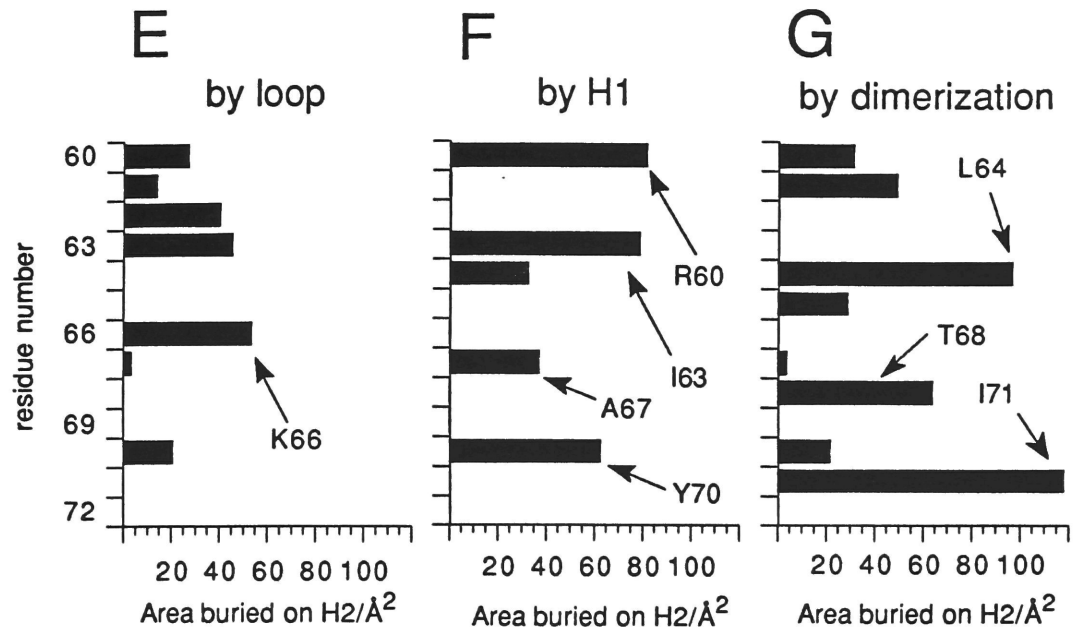
The reduction in solvent accessible surface areas resulting from dimerization and DNA binding are shown on a per-residue basis in Fig. 18A. It is clear that dimerization results from the interaction of residues in the H1, H2 and zipper regions, while the amino-acid residues which participate in DNA-binding are present predominantly in the basic region, with the exception of three residues at the end of the loop and the beginning of the H2 helix.

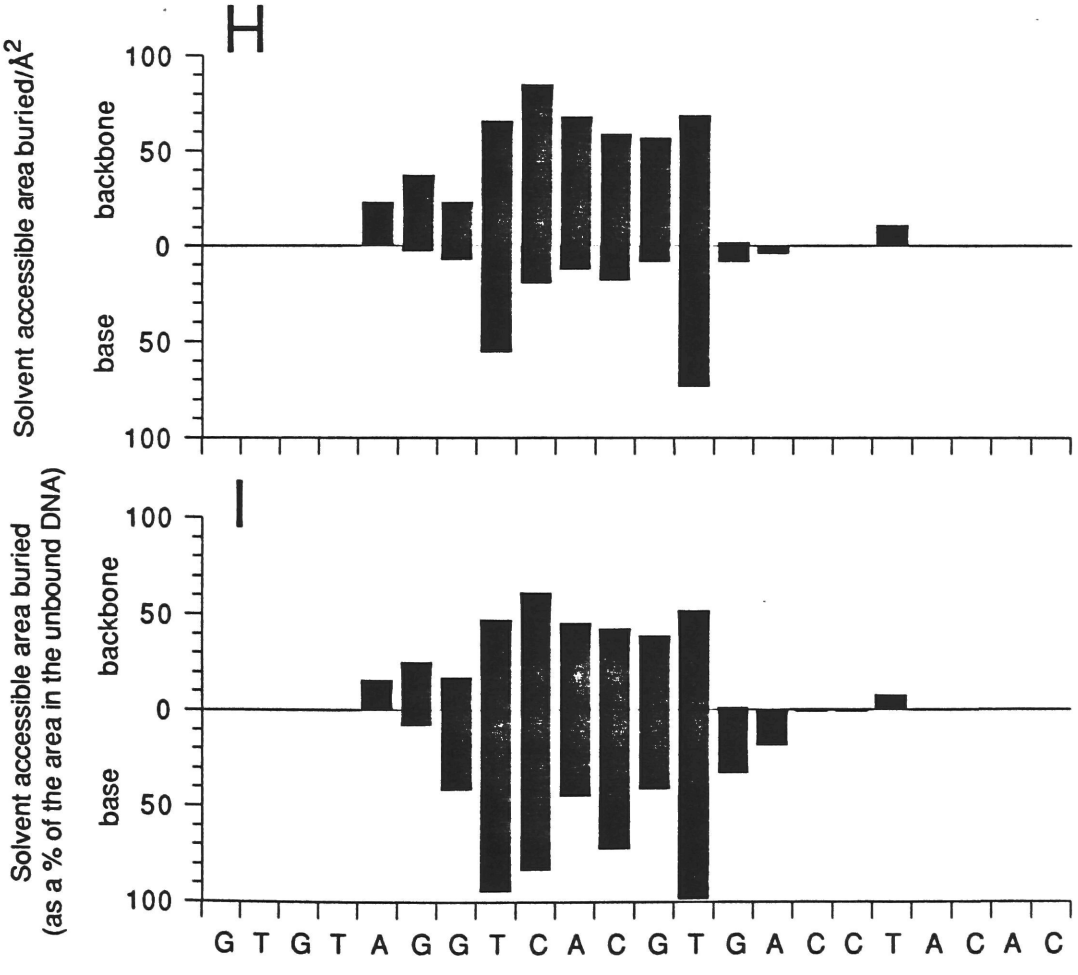
The reduction in solvent-accessible surface area of H1 resulting from tertiary packing interactions are shown in Fig. 18B, C, and D. The corresponding reductions for H2 are shown in Fig. 18E, F, and G. As expected from the overall topology of the structure, the loop packs against the upper (in the canonical view of Fig. 17) end of H1 and the lower end of H2. Loop residues are not involved in dimerization. H1 and H2 from the same protomer have an interaction surface comprising about five residues each. A slightly larger number of residues from both of these regions participates in forming the dimer interface. The coiled-coil zipper region only packs against its dimerization partner.

DNA Structure The reduction in solvent-accessible surface area on a strand of DNA by the binding of a monomer of Max is shown in Figs. 18H and 18I for the base and backbone atoms of each base. It can be seen that the majority of the DNA-protein contacts take place near the recognition element CACGTG, and predominantly on the 5' half of the DNA. Analysis of the DNA stereochemistry employing CURVES and a global helix axis

Figure 18. Solvent accessible surface areas buried upon interaction of various parts of the refined $(Max22\text{-}113)_2$-DNA complex model. Accessible surface areas were calculated using a water probe radius of 1.4 Å. (**A**) Solvent accessible surface areas buried per amino acid residue upon dimerization and DNA-binding. Some residues making important interactions are indicated. (**B**) Solvent accessible surface areas buried per amino acid residue on the H1 segment of the protein upon interaction with the loop region or, (**C**) the H2 region, or (**D**) upon dimerization, as in (A). (**E**) Solvent accessible surface areas buried per amino acid residue on the H2 segment of the protein upon interaction with the loop region or, (**F**) the H1 region, or (**G**) upon dimerization, as in (A). (**H**) Solvent accessible surface areas buried upon (imaginary) binding of a protomer of Max22-113 to one strand of DNA for base and phosphodiester backbone (including deoxyriboses) atoms of the DNA. (**I**) Same as (H) but expressed as a percentage of the solvent accessible surface area in the unliganded DNA.


A
Solvent accessible area buried/Å² per monomer
on DNA binding
on dimerization
90
60
30
0
30
60
90
H28
L31
K40
R35
L46
R60
L64
I71
H81
L88
L102
22 27 32 37 42 47 52 57 62 67 72 77 82 87 92 97 102107
residue number
B
by loop
C
by H2
D
by dimerization
residue number
36
39
42
45
48
51
R47
P51
F43
L46
I39
L46
S49
20 40 60 80
Area buried on H1/Å²


E
by loop
F
by H1
G
by dimerization
residue number
60
63
66
69
72
K66
R60
I63
A67
Y70
L64
T68
I71
20 40 60 80 100
Area buried on H2/Å2

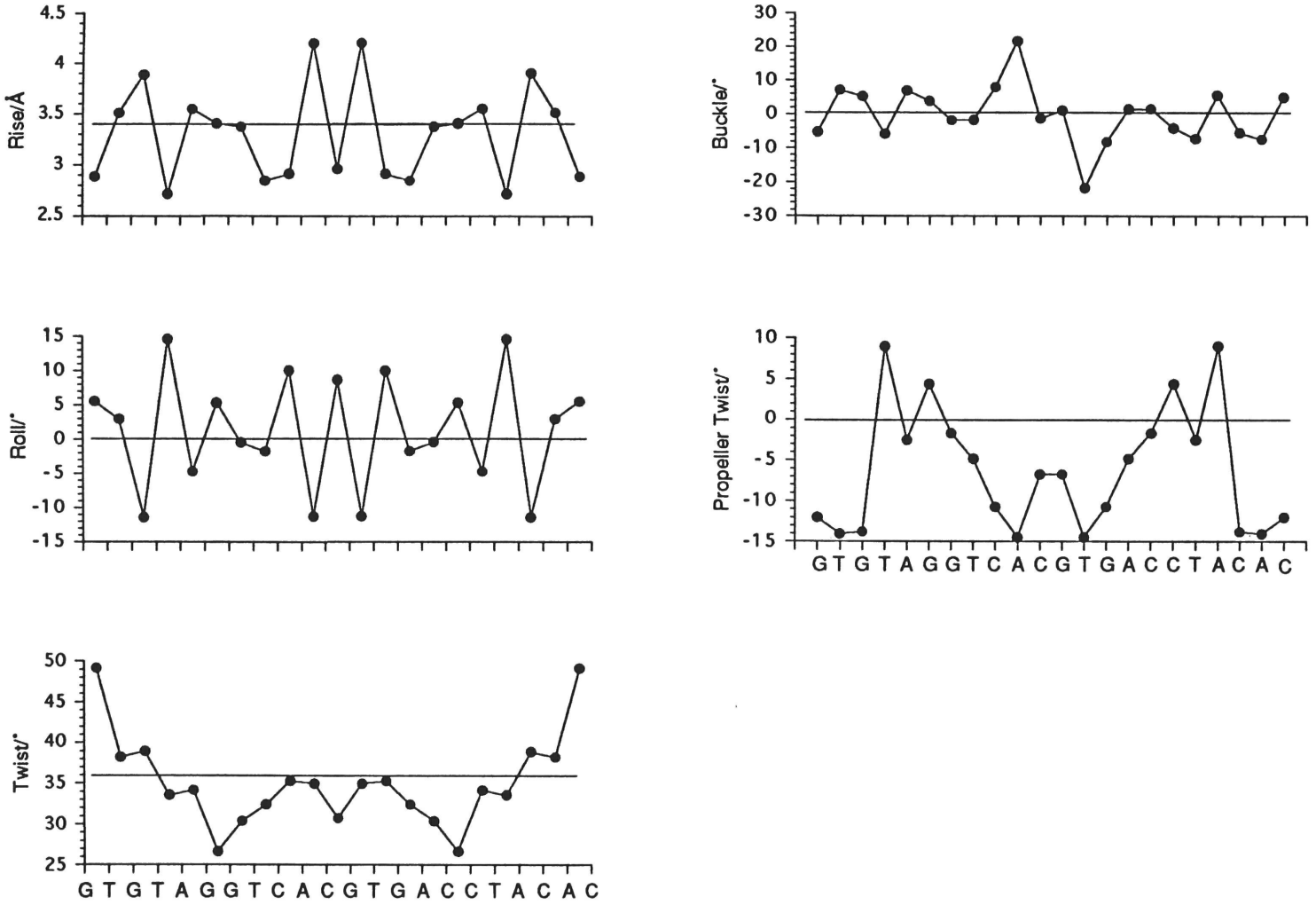

H
I
Solvent accessible area buried/Å2
Solvent accessible area buried
(as a % of the area in the unbound DNA)
backbone
base
100
50
0
G T G T A G G T C A C G T G A C C T A C A C

reveals only small deviations from canonical B-form DNA. The mean distance between phosphorus atoms is 6.7 Å (standard deviation 0.4 Å; comparable with the 7 Å distance expected in "standard" B-form DNA with all 2'-endo puckers, Saenger, 1984, p. 229), the average twist is 35.2°, implying 10.2 bp per turn, and the mean rise per base is 3.32 Å (for comparison, values of twist and rise range from 30.0 to 32.7° and 3.03 to 3.37 Å, respectively, for "typical" B-form DNA; Saenger, 1984, p. 229). Rise, roll, and twist per base pair and buckle and propeller twist per base are plotted in Fig. 19. Ignoring end effects, the most striking feature is the pronounced roll, buckle and propeller twist of the A:T base pair in the recognition element.

Max Monomer Structure The polypeptide backbone adopts α-helix hydrogen bonding starting with Arginine 25, in the basic region, and deviates significantly from it at Serine 49, near the end of H1. The phylogenetically conserved hydrophobic amino acid following this residue (Valine 50 in Max) packs against a conserved tyrosine (Tyrosine 70) in the H2/Z α-helix. The conserved proline at position 51 provokes a turn in the backbone, starting the eight amino acid loop that connects the two α-helical segments. The 43 residue long H2/Z α-helix begins at the conserved basic residue Arginine 60 and extends to Leucine 102, the last leucine of the conserved heptad repeat. Of the remaining eleven residues, five have been built as random coil while the C-terminal six amino acids are not visible in the electron density map.

Parallel Four-Helix Bundle The parallel, left-handed, four-helix bundle, formed by b/H1 and H2/Z derived from each member of the symmetric homodimer creates the hydrophobic core of the Max-DNA complex. Fig. 20A shows the interactions between the H1 and H2 regions of a monomer that create half of the four-helix bundle. Isoleucine 39 and Phenylalanine 43 from H1 interact with Arginine 60, the first residue of H2. Above

Figure 19. Some stereochemical parameters of the DNA in complex with Max. The DNA of the refined $(Max22\text{-}113)_2$-DNA complex model was analyzed using the program CURVES (Lavery and Sklenar, 1989) and a global helix axis. The parameter plotted is on the ordinates of the graphs; the DNA sequence is in the abscissa.


Rise/Å
4.5
4
3.5
3
2.5
Buckle/°
30
20
10
0
-10
-20
-30
Roll/°
15
10
5
0
-5
-10
-15
Propeller Twist/°
10
5
0
-5
-10
-15
Twist/°
50
45
40
35
30
25
G T G T A G G T C A C G T G A C C T A C A C

these residues lie Leucine 46 and Valine 50 of H1 and Isoleucine 63, Leucine 64, Alanine 67 and Tyrosine 70 of H2. Tyrosine 70 packs against Valine 50 and Proline 51.

The dimerization interface can be thought of as being the result of packing the two H2 α-helices together in a parallel orientation and then overlaying the two H1 α-helices on either side. Fig. 20B shows the two H2 regions that participate in forming the four-helix bundle. Isoleucine 63 and Leucine 64 from both protomers pack closely together as do residues more C-terminal such as Isoleucine 71 and Methionine 74. The central portion of this H2 dimer is somewhat open, being occupied only by the phylogenetically conserved Alanine 67 from both protomers. H1 packs into this relatively open section of the H2 pair, inserting Leucine 46 and Valine 50 from H1 into the space left vacant, as it were, by Alanine 67 (Fig. 20C). Leucine 46 is the only residue from the H1 segment which comes to within van der Waals contact distance with its dyad-related Leucine 46' (3.5 Å in the refined model). For comparison, Phenylalanine 43 and Phenylalanine 43' are 4.6 Å apart at the point of closest approach.

Leucine Zipper Beyond Methionine 74 lies the coiled-coil formed by the two zipper regions (Fig. 20D), an extension of the H2 coiled-coil. In addition to the Leucine-Leucine and Isoleucine-Isoleucine interactions seen at positions 88, 95, 102 (Leucine) and 85 (Isoleucine), there is an Asparagine-Asparagine interaction at position 78, a Histidine-Glutamine-Histidine-Glutamine tetrad at positions 81 and 82, and an Asparagine-Glutamine-Asparagine, Glutamine tetrad at positions 91 and 92.

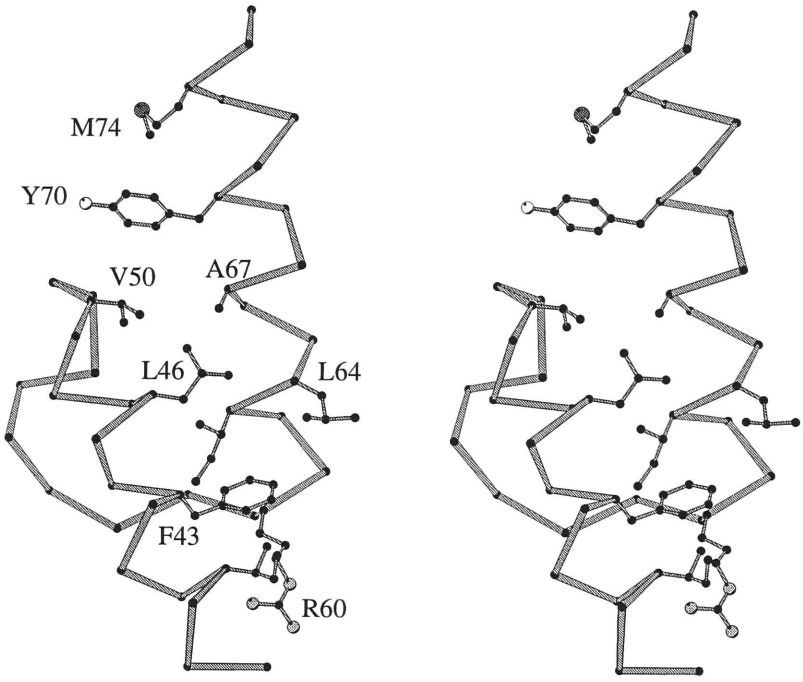
Loop The structure of the (Max22-113)-DNA complex fixes the distance between the last α-carbon of H1 (Proline 51) and the first α-carbon of H2 (Arginine 60) at 15Å. The loop region (Fig. 20D) is well defined in the electron density maps, presumably because its conformation is stabilized by interactions between Nη of Arginine 47 and the carbonyl

oxygen of Lysine 57, interactions of this Lysine and Serine 59 with the DNA backbone, as well as by crystal packing interactions (see below). In addition, Leucine 53 in the loop region packs against the external edge of the hydrophobic core of the four-helix bundle, constituted by the aliphatic portions of Arginine 47 from H1 and Lysine 66 from H2 as well as by Valine 50 at the end of H1 (Fig. 20E).

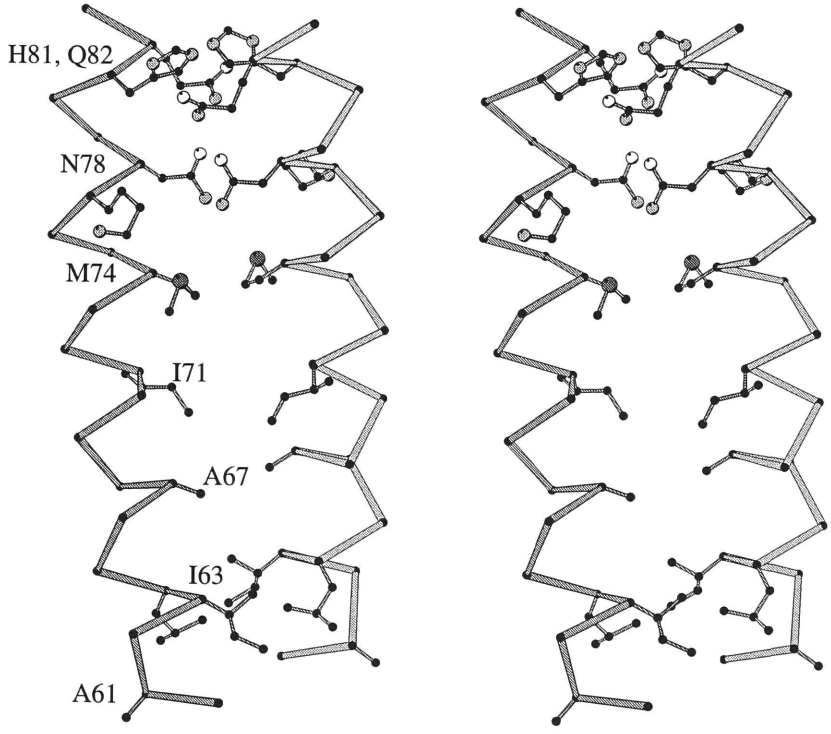
DNA-Protein Interface Three distinct portions of Max22-113 interact with DNA (Figs. 13, 17, 18H, 20E, 20F, 21). First, three amino acid residues from the α-helical basic region make direct contacts with DNA bases in the recognition sequence CACGTG: Histidine 28 with G3R, Glutamate 32 with C3L', A2L' (where the prime denotes the dyad-related molecule) and T2R, and Arginine 36 with G1R. In addition Arginine 36 appears to stabilize the position of Glutamate 32 by hydrogen bonding both to it and to the phosphate backbone (Figs. 13B, 21). The basic region also makes a large number of phosphate contacts, which span the entire backbone of the recognition element. The side-chains making phosphate contacts are Arginine 25, Asparagine 29, Arginine 33, Arginine 35, Arginine 36 from the basic region and Lysine 40 from H1. In addition Leucine 31 which is located within the major groove of the DNA suffers a considerable loss of solvent-accessibility upon association of the protein with DNA (Figs. 20F, 21). Second, Lysine 57 in the loop region of Max stabilizes the meandering path of the loop, which extends across the adjacent minor groove. This structural feature allows the lysine side chain to straddle the minor groove, making a salt bridge with the phosphodiester backbone on its opposite side (Fig. 20E). Third, Arginine 60, which is located at the start of H2, makes a side chain contact with the phosphate of C3 (Fig. 20C) and a main chain amide contact with the phosphate of A2. In addition, Serine 59 at the C-terminal end of the loop makes a side-chain to phosphate (of T4) contact. The protein-DNA contacts will be further analyzed and put in the context of mutagenesis results on various helix-loop-helix proteins in Chapter 4.

Figure 20. Parallel-eye stereo representations of portions of the $(Max22\text{-}113)_2$-DNA complex. Figures were prepared using MOLSCRIPT (Kraulis, 1991). The C-terminus of the protein always lies towards the top of the page. **(A)** The HLH region of a single protomer. Residues contributing to the H1:H2 interface are shown as ball-and-stick figures. **(B)** The coiled-coil formed by H2 regions from two protomers. The first α-carbon shown (at the bottom of the figure) corresponds to Arginine 60. Buried side-chains are represented in ball-and-stick representation. Note the "hole" left by the Alanine 67 residues from the two molecules. The zipper consensus starts with Methionine 67. **(C)** The result of positioning the H1 region from one protomer onto the H2:H2 coiled coil in the same orientation as (B). The loop region α-carbons are connected with dotted lines. Note how Leucine 46 and Valine 50 fit into the hole left by Alanines 67 from H2. The interaction of Phenylalanine 43 with Arginine 60 and Isoleucine 63 from the same protomer and Leucine 64 from the other protomer is also apparent. **(D)** Zipper region coiled coil. Note the interfacial Histidine 81-Glutamine 82 and the Glutamine 91-Asparagine 92 tetrads. Four α-carbons beyond residue 103 are connected with dotted lines. **(E)** Interactions of the loop with the upper edge of the four-helix bundle and the phosphodiester backbone. Note how Leucine 53 inserts into a cavity on the outer surface of the hydrophobic core, and packs against the aliphatic portions of Arginine 47 and Lysine 66, as well as Valine 50. DNA backbone atoms bordering the minor groove are shown. Note how Lysine 57 straddles the groove to make a single terminal amine-phosphate oxygen contact. Also apparent are the phosphate contacts made by Serine 59 and Arginine 60. Arginine 36 from the basic region, which lies deep in the major groove, is shown for reference. **(F)** Interactions between the basic region and the phosphodiester backbone. For comparison with figure 21, side chains which make base contacts are also shown. The bottom of the H2 helix of the dyad related protomer is also shown. Lysine 24 and Leucine 31 do not make any contacts with the DNA.

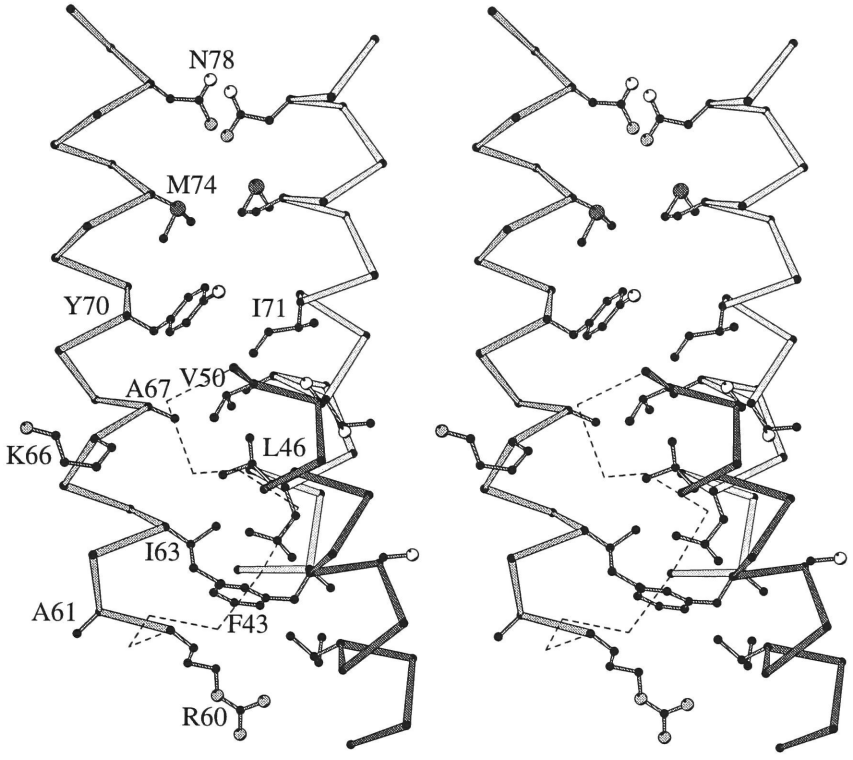
A



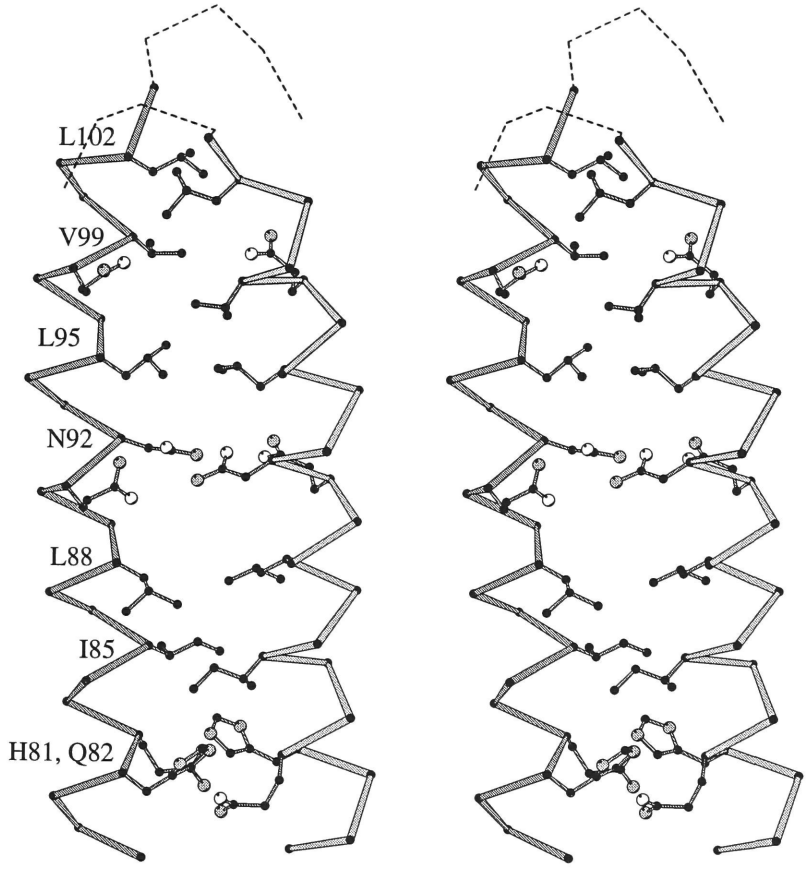
B



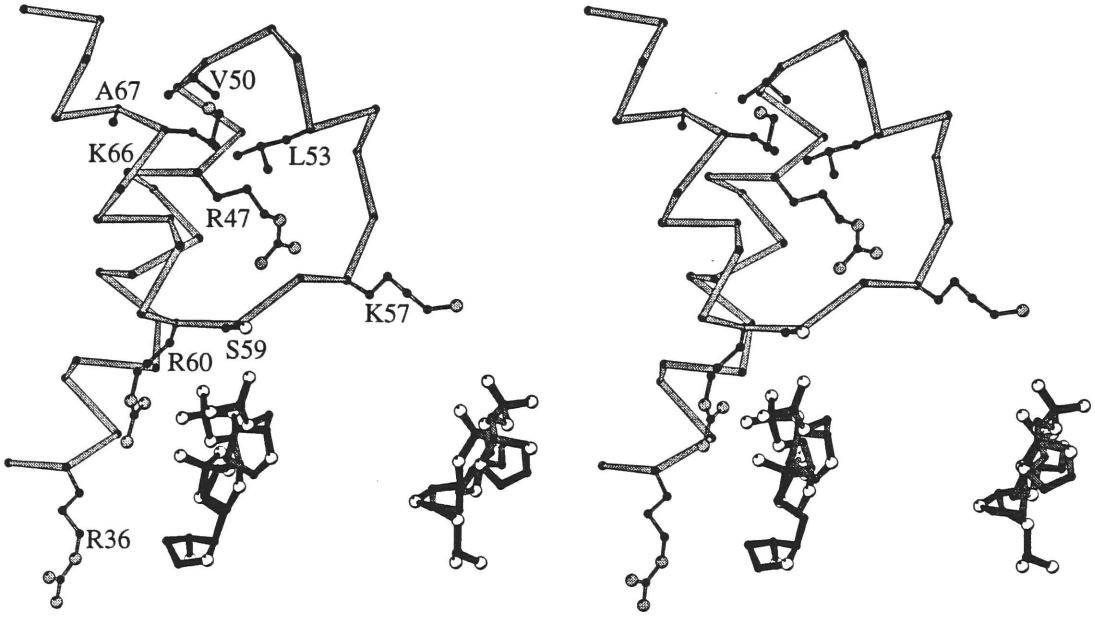
C


N78
M74
Y70
I71
A67
V50
K66
L46
I63
A61
F43
R60

D


L102
V99
L95
N92
L88
I85
H81, Q82

E


A67
V50
K66
L53
R47
K57
S59
R60
R36

F

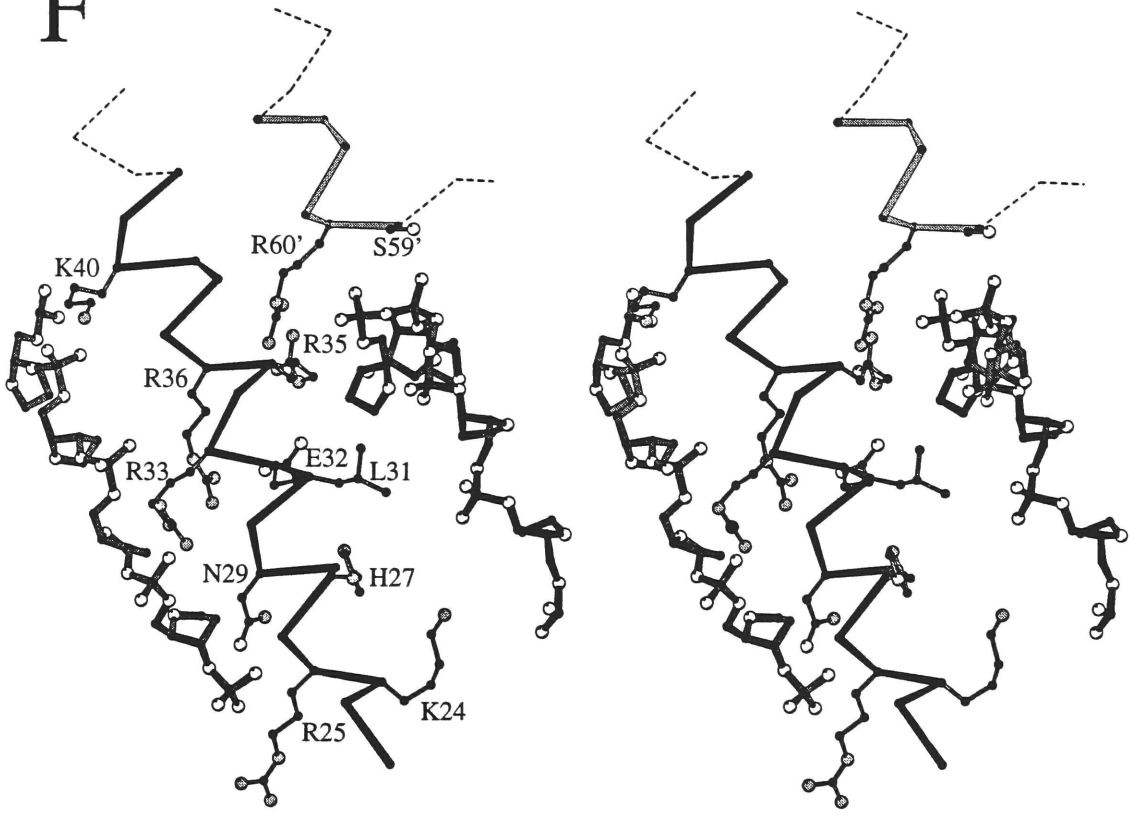
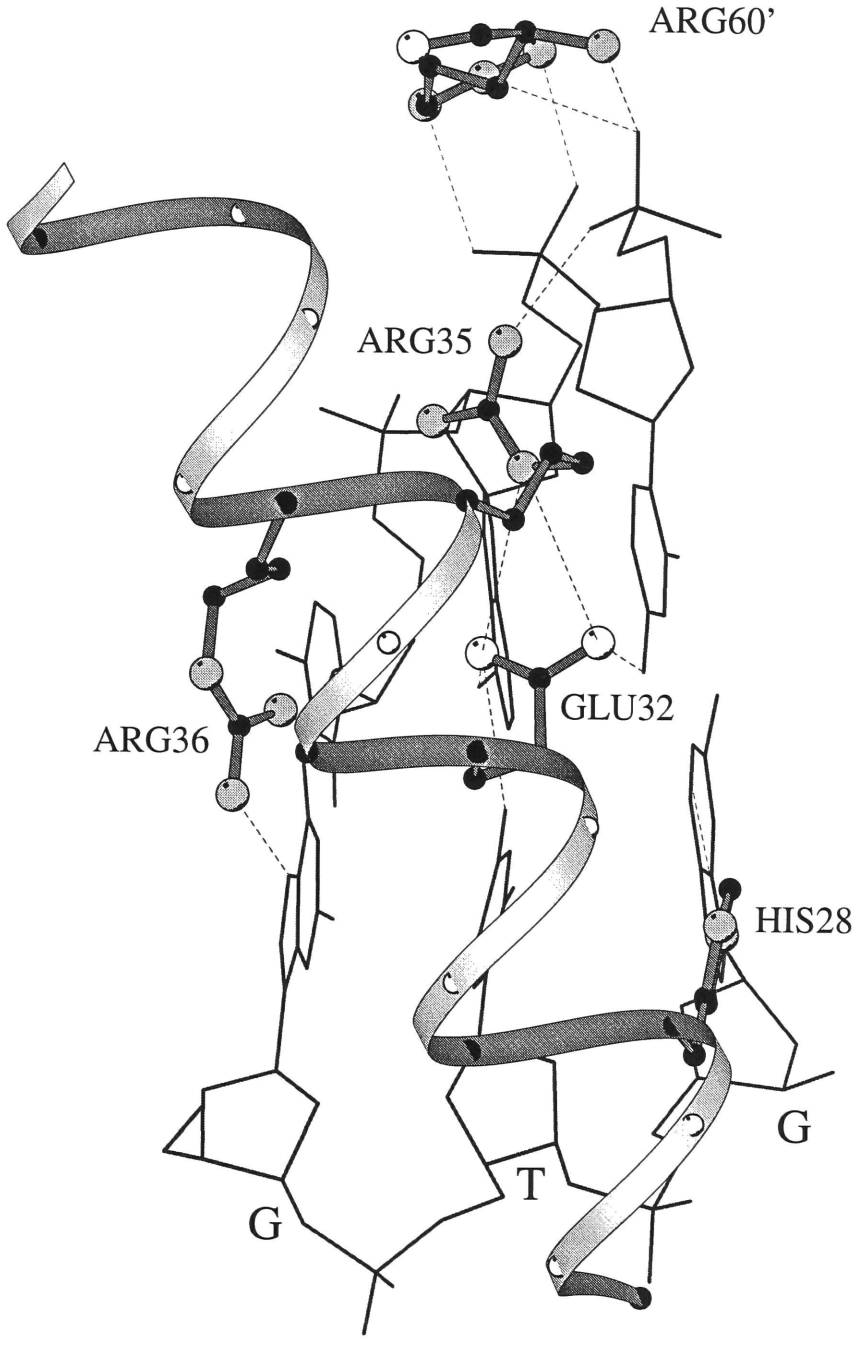

R60'
S59'
K40
R35
R36
R33
E32
L31
N29
H27
K24
R25

Figure 21. Max basic region-E-box interactions. The DNA is represented in thin lines, and the path of the protein backbone as a ribbon, with α-carbons drawn as spheres. α-carbons of protein side-chains making contacts with the phosphate backbone are colored black, those not interacting with the DNA white. Protein side-chains interacting with DNA bases are shown as ball-and stick figures. Figure prepared using MOLSCRIPT (Kraulis, 1991).


ARG60'
ARG35
GLU32
ARG36
HIS28
G
T
G

Crystal Packing In addition to the specific protein-DNA interactions enumerated above, the Max (22-113)-MLP22 cocrystal shows three sets of DNA-protein interactions (Fig. 13C). First, the blunt-ended DNA packs against the upper portion of the loop and the α-helical region corresponding to the end of H2 and the beginning of Z. This contact involves mostly phosphate backbone to basic amino acid side-chain contacts, as would be expected from the two kinds of blunt ends present in the crystal due to static disorder. Second, the uppermost portion of the leucine zipper packs against the major groove edge of a symmetry related DNA duplex. Third, the back side of the basic region α-helix packs against the backbone of a symmetry related DNA. The three kinds of protein-DNA interactions appear to stabilize the crystal lattice. A view of the molecular packing looking down the crystallographic six-fold screw axis is shown in Fig. 13C. It can be readily appreciated that the DNA does not stack end-to-end to form a pseudo-continuous helix as seen in most DNA-protein cocrystals (Anderson *et al.*, 1984; see also Brennan *et al.*, 1986).

3.4 Structure of a USF b/HLH Dimer Bound to DNA

Structure determination The b/HLH construct of USF was cocrystallized with the 21-mer oligonucleotide duplex derived from the adenovirus major late promoter sequence shown in Fig. 4A as described in Section 2.8. The crystals (Fig. 11C) were orthorhombic (space group $P2_12_12_1$, determined ultimately by molecular replacement; a = 136.6Å, b = 54.7Å, c = 44.4Å; one $(b/HLH)_2$-DNA complex per asymmetric unit). At 4°C, the crystals were very susceptible to radiation damage, making data collection impossible. Despite considerable persistence, suitable flash freezing conditions could not be established. Oscillation photographs were collected at -20 °C, from one crystal using CuKα X-radiation with a Rigaku RAXIS-IIc area detector, from three crystals with 0.91Å X-radiation and imaging plates at beam-line F1 of the Cornell High Energy Synchrotron

Source (CHESS), and from three crystals with 0.95 Å X-radiation and imaging plates at the wiggler beam-line X25 of the National Synchrotron Light Source (NSLS), Brookhaven National Laboratory. Crystals were mounted in capillaries, usually with the *a* axis nearly parallel to the spindle. Diffraction data were reduced to structure factor amplitudes and scaled using the programs DENZO and SCALEPACK. Fresh crystals diffracted isotropically to minimum Bragg spacings of $1/2.6Å^{-1}$ (Fig. 11D), but severe radiation sensitivity limited useful data to 3.5Å in the laboratory and 2.9Å at CHESS and NSLS. Ultimately, data from all seven crystals were merged. Data statistics appear on Table 8.

The structure was solved by molecular replacement using a dimer of amino acids 22 through 80 and the central 10 bp (*i.e.* a total of 20 bp) of the refined Max-DNA structure as a search model. A rotation search between 10.0 and 4.5 Å, followed by Patterson correlation refinement with X-PLOR yielded a solution 4.8σ above mean peak height with Euler angles θ_1, θ_2, θ_3 equal to 272.1°, 72.5°, and 177.9°, respectively. This solution was employed in a translation search with X-PLOR, resulting in a family of symmetry-related solutions 3.2σ above mean peak level. The model was modified at this stage by incorporating the correct DNA sequence and the overhanging nucleotides [for which electron density was present in $(2|F_{observed}|-|F_{calculated}|)$ electron density maps]. All side-chains were stripped, and the polyalanine model without the loop region with the full 21 bp DNA was subjected to positional refinement to yield an *R*-factor of 36.7% from 10.0 to 3.1Å. Side-chains, loop, and N-terminal residues were gradually added to the model by inspecting omit and annealed-omit maps. Manual rebuilding employing O and positional refinement brought the *R*-factor to 28.9%. At this stage, the DNA sequence was inverted, resulting in a decrease of the *R*-factor by 1.5%; this model had an *R*-factor of 27.1% between 6.0 and 3.1Å, using an overall isotropic temperature factor of $17.0Å^2$. The model had an *R*-factor of 25.1% for the same resolution range using a *B*-factor model consisting of six groups. The values of the *B*-factors ranged from 21.8 $Å^2$ to 56.6 $Å^2$.

Table 8. Statistics of merged USF $(bHLH)_2$-DNA diffraction data

Resolution range (Å)	15.0 - 2.9	3.0 -2.9
Reflections, observed	45390	
Reflections, unique	6038	375
Average redundancy	2.6	1.5
Data coverage (%)	77.2	49.3
R_{sym} ($\|F\|>0$) (%)*	7.6	32.5
$\langle I/\sigma(I)\rangle$ (%)	14.4	3.18

* $R_{sym} = \Sigma\, |I-\langle I\rangle|/\Sigma\, I$, whre I is the observed intensity and $\langle I\rangle$ is the average intensity obtained from multiple observations of symmetry related reflections.

Table 9. Crystallographic refinement statistics of the $(b/HLH)_2$-DNA complex

Resolution range (Å)	6.0-2.9
R factor (%)	23.6
Reflections ($\|F\| > 1\sigma\|F\|$)	5096
Total number of non-hydrogen atoms	1923
R.m.s. bond lengths (Å)*	0.019
R.m.s. bond angles (degrees)*	3.05
R.m.s B-factors bonded atoms (Å^2)*	2.7
Average B-factor protein main chain (Å^2)†	47.5
Average B-factor protein side chains (Å^2)†	44.0
Average B-factor DNA bases (Å^2)†	24.0
Average B-factor DNA backbone (Å^2)†	36.6

* R.m.s. bond lengths and r.m.s. bond angles are the respective root-mean-square deviations from ideal values; r.m.s. B-factor is the root-mean-square deviation between the thermal parameters of coavalently bonded atomic pairs.

† Average B-factors are arithmetic means.

Further manual rebuilding, positional refinement, phase extension to 2.9Å, and refinement of tightly-restrained individual isotropic temperature factors resulted in the current model. Refinement statistics for this model are presented in Table 9, and a portion of the final electron density map is shown in Fig. 22E. Given the marginal completeness of the diffraction data, it is not surprising that the quality of the current model is modest. A Luzzati plot (not shown) indicated that the precision of the coordinates is approximately 0.4 Å. When the ϕ φ angles of the model were analyzed with PROCHECK, 4.2% (5 out of 130) were found to lie in disallowed regions of the Ramachandran plot.

Comparison with the Max cocrystal structure The (b/HLH)$_2$-DNA complex folds into a parallel, left-handed four-helix bundle, which is topologically identical to the structure of Max (Fig. 22A, compare with Fig. 17). Comparison of the two three-dimensional structures suggests that the basic and loop regions of the USF b/HLH construct are remarkably plastic. Several differences between the Max and USF structures represent distortions of the polypeptide chain that can be attributed to crystal packing interactions (Fig. 22E). The N-terminal 10 amino acids of molecule 2 of USF (yellow) have been pulled down from its cognate DNA by a lattice contact with a neighboring DNA duplex. The top 4 residues of helix 1 of USF (H1) and the loop region of molecule 1 (red) have been stretched upwards by another crystal packing interaction. Finally, the C-terminal 4 residues of both molecules of USF deviate from α-helical geometry because of lattice contacts with a pair of symmetry-related complexes. The undistorted portions of the USF complex show that the basic (b) and H1 regions (amino acids 199 through 225) form an uninterrupted α-helix, as do amino acids 243 through 256 of H2. The hydrophobic core of the four-helix bundle closely resembles that of Max, and the last conserved hydrophobic position of H2, Leucine 254 (equivalent to Methionine 74 in Max), packs against its dimer mate in standard coiled-coil fashion. The b/HLH expression construct encoded another six amino acids beyond Leucine 254, of which the last four appear as random coil in the

structure. Presumably these residues do not adopt the left-handed coiled coil of right-handed α-helices seen in the Max (b/HLH/Z)$_2$-DNA structure because the first leucine of the leucine zipper (Z) or heptad repeat region is not present in my b/HLH construct of USF.

The conformation of the DNA complexed with b/HLH is not systematically different from B-form DNA. The mean rise per base pair is 3.32Å and the average helical twist is 32.9°, implying 10.9 bp/turn. The DNA stacks 5' to 3' in the crystal, producing a pseudo-continuous, B-form double helix which is stabilized by Watson-Crick base pairing of the overhanging C and G. The pronounced buckle and propeller twist of the A:T base-pairs of the E-box observed in the Max cocrystal structure was also present here.

In the orthorhombic crystal form examined here, the two basic regions of the b/HLH dimer lie in different crystalline environments and the basic region of molecule 2 (yellow) appears to be distorted by lattice contacts (Figs. 22A, 22C). Comparing interactions made by the two polypeptide chains with DNA, stronger, presumably specific, interactions can be distinguished from weaker, or secondary, interactions. Two critical side chain-base contacts are made by both basic regions. Arginine 212 (homologous to Arginine 36 in Max) contacts N7 of the guanine adjacent to the palindrome's dyad axis through its η1 nitrogen. Glutamate 208 (homologous to Glutamate 32 in Max) makes contact with the N4 of the outer C of the palindrome. In addition to these amino acid-base contacts, backbone contacts involving Asparagine 205, Arginine 209, and Arginine 211 are seen in positions equivalent to those seen in the cocrystal structure of the complex of Max with CACGTG. In addition, Nε2 of Histidine 204 of molecule 1 is located 3.8Å away from O6 of the outer G (G3). In the Max structure, the equivalent Histidine 28 is located a similar distance away from N7 of the same G.

Comparison of the undistorted basic region of molecule 1 of USF with the corresponding region of Max (Fig. 22D) demonstrates a high degree of structural similarity, with most equivalent residues lying in approximately equal conformations. Nonetheless, the orientations of the helices relative to the major groove of DNA appears to differ considerably. This may result, for instance, in the difference in contacts made by the histidine residues, mentioned above.

Conformational differences between the two b/HLH basic regions of USF underscore the relative plasticity of this portion of the motif. The Max-like basic region of molecule 1 participates in a dense network of side chain-DNA contacts involving Glutamine 203, Histidine 204, Arginine 210, Aspartate 213 and Asparagine 216, in addition to those mentioned above. The distorted basic region of molecule 2 makes substantially fewer contacts with DNA. Arginine 210 and Asparagine 216 make the same contacts as in molecule 1. Because of polypeptide backbone distortion, Asparagine 205 contacts a phosphate one nucleotide removed from that contacted by both the undistorted molecule 1 and the corresponding residue in the Max-DNA complex structure. Finally, the side chains of Arginine 200 of molecule 1 and Glutamine 203 of the unwound basic region of molecule 2 make contacts with bases outside the central, palindromic CACGTG element.

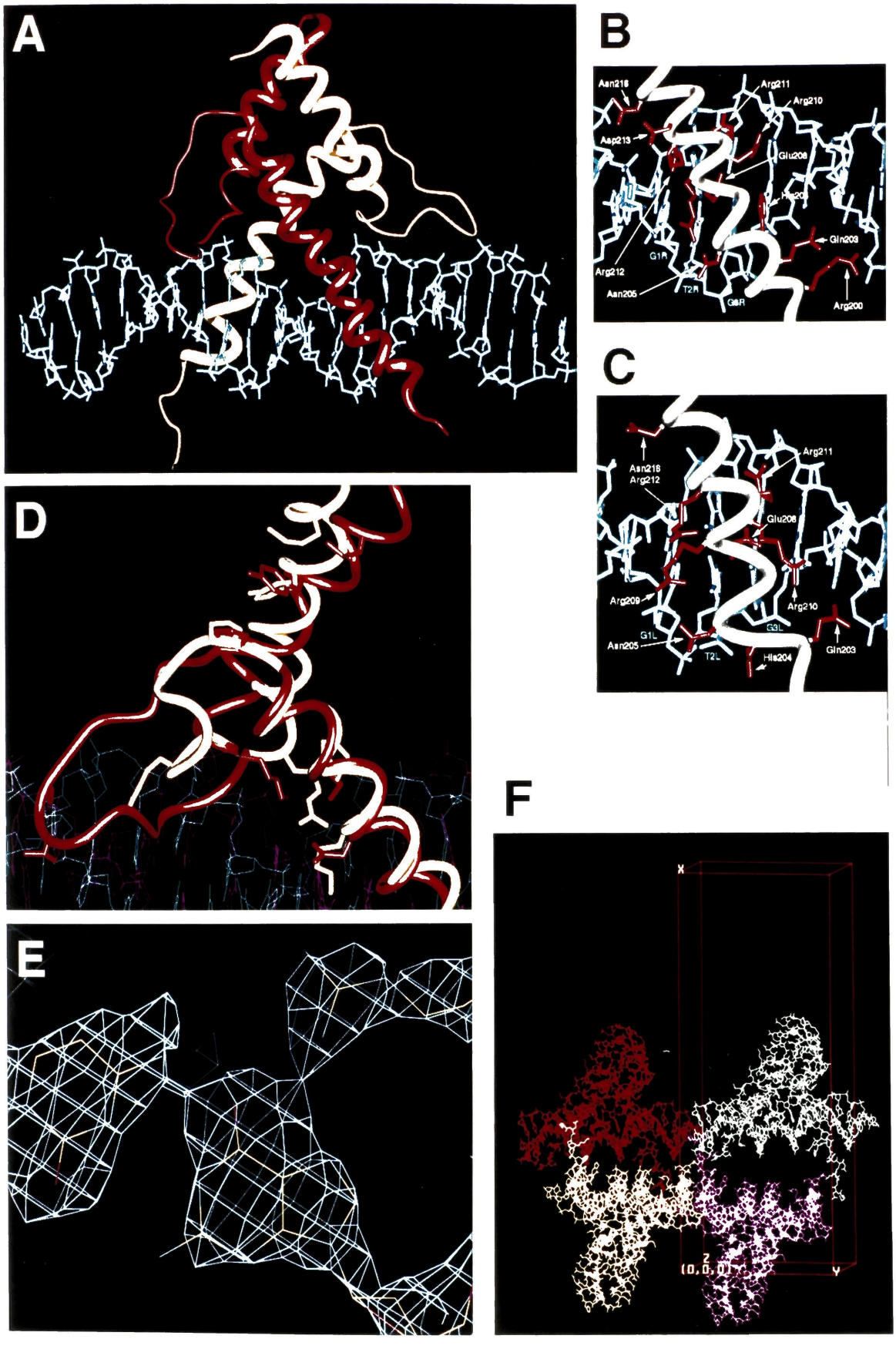
Loop As in the Max-DNA complex structure, both the loops and the four-helix bundle of USF interact with DNA. The USF loop region consists of 12 amino acids, four residues longer than in Max. As a result of this added length, the loop of molecule 2 traverses the adjacent minor groove and makes two phosphate contacts (Serine 233 and Threonine 234; Fig. 22D) and a contact with a sugar oxygen within the minor groove (Glutamine 238 Nε to C4R O4'; 2.8Å). In contrast, the eight-residue Max loop makes a sole lysine-phosphate contact (Lysine 57; Fig. 20E). The loop of molecule 1 (red) adopts a different conformation than either the Max loop or the loop of molecule 2 of USF (Fig. 4A, 4D). In

this case, the loop is involved in a lattice contact. The backbone amide of the conserved basic residue at the beginning of H2, Lysine 240, packs against the phosphate backbone, as does the equivalent residue of Max (Arginine 60).

Helix-Loop-Helix Packing Three-dimensional alignment of the HLH regions of USF and Max (Fig. 22D) shows that the crossing angle between H1 and H2 is substantially different in the two structures. As mentioned above, the hydrophobic cores of the four-helix bundles are similar, and a number of interactions which presumably affix H1 to H2 are conserved. For instance, Isoleucine 219 from H1 packs against Lysine 240 at the beginning of H2 in the USF structure as does Proline 227 at the end of H1 against Tyrosine 250 on the outer face of H2, interactions analogous to the Phenylalanine 43-Arginine 60 and Proline 51-Tyrosine 70 contacts seen in Max. The differing orientations of H2 will be further discussed in Section 4.1.

Crystal packing Diffraction pictures taken from USF b/HLH-MLP21 cocrystals show one pair of intense meridional reflections if the X-rays are perpendicular to the longest unit cell edge, and *two* pairs, if the X-rays are parallel to the same edge (Fig. 13E). The packing of the DNA-protein complex in the crystal (Fig. 22F) is such that the pseudo-continuous DNA-helix resulting from stacking of the synthetic duplexes and base-pairing of the G:C overhangs runs diagonally across the b x c plane of the unit cell in one asymmetric unit, and because of the two-fold screw axis, it runs also diagonally but 90° away from the previous DNA in the next asymmetric unit, thus accounting for the striking diffraction pattern. The protein moieties of the complex sit sandwiched between these alternating layers of diagonal DNA duplexes, making the extensive set of lattice contacts described above. In addition to the intense meridional reflections, strong diffuse scattering, both as halos around the Bragg spots and as streaks between Bragg spots were noticeable in diffraction pictures from USF b/HLH-MLP21 cocrystals (Fig. 11D). Such diffuse

Figure 22. Some salient features of the USF (b/HLH)$_2$-DNA structure. (**A**) Overall view of the complex. Helical regions are represented by thick tubes; irregular secondary structure elements by thin tubes. Molecule 1 and molecule 2 are colored red and yellow, respectively. The N-termini are at the bottom of the figure. (**B**) View of the basic region of molecule 1 interacting with DNA. The path of the protein backbone is shown as a white tube; side chains which interact with DNA as red stick figures. (**C**) View of the basic region of molecule 2 as in (B). Note that the first turn of polypeptide backbone is not α-helical. (**D**) Alignment of the H1, loop and H2 regions of molecule 2 of USF (in red, DNA in blue) with the corresponding regions of Max (in yellow, DNA in magenta). The superposition was achieved by fitting the α-carbon coordinates of USF residues 216 through 225 with the corresponding atoms of Max. The rmsd of the 10 pairs of atoms is 0.63 Å. (**E**) Final electron density of USF (b/HLH)$_2$-DNA. The $(2|F_{observed}| - |F_{calculated}|)$ Fourier synthesis was contoured at 1.4 σ. The C:G base pair immediately adjacent to the dyad axis of the E-box and the side chain of Arginine 212 which interacts with it are shown. (**F**) Crystal packing of USF (b/HLH)$_2$-DNA. The macromolecular contents of four asymmetric units are shown in different colors. Note how the two pseudo-continuous DNA helices are almost orthogonal to each other. Compare with Figs. 11E and 13C. The boundaries of one unit cell are shown [the origin is marked (0,0,0)].


A
B
Asn218
Arg211
Arg210
Asp213
Glu208
Gln209
G1R
Arg212
T2R
Asn205
G3R
Arg200
C
Asn218
Arg212
Arg211
Glu208
Arg209
Arg210
G3L
G1L
Asn205
T2L
His204
Gln203
D
E
F
X
Z
(0,0,0)
Y

features were present, but to a lesser extent, in the Max b/HLH/Z-MLP22 cocrystal diffraction pictures (Fig. 11B).

3.5 Variations on the HLH Theme: Work in Progress

Ongoing structural and physico-chemical investigation of helix-loop-helix proteins aims to achieve a refined understanding of dimerization and DNA-binding specificity. One line of research is concerned with the structural and biochemical characterization of b/HLH and b/HLH/Z proteins which interact with atypical DNA sequences. Understanding how the b/HLH framework can be adapted for the recognition of non-E-box sequences is expected to shed light also on conventional E-box recognition. Investigation of dimerization specificity requires preparation of heterodimers. This kind of work is complicated by the difficulty in avoiding formation of homodimers when two different b/HLH or b/HLH/Z proteins are mixed together. We have adopted a chemical solution to this problem.

Atypical b/HLH:DNA Interactions I: USF and the LCR Bresnick and Felsenfeld (1993) reported that USF modulates globin transcription by binding to the atypical E-box caCCtg present in the locus control region. Fig. 23A shows that USF b/HLH undergoes a folding transition that is intermediate to that resulting from challenge with the MLP E-box and nsDNA when challenged with this element. USF b/HLH crystallized under conditions similar to those employed with the MLP 21-mer when mixed with a 21-mer DNA in which the MLP E-box had been substituted with the LCR element. In order to obtain more frequent nucleation, the terminal residues were substituted to the residues present in the oligonucleotide shown in Fig. 4B Since the b/HLH-LCR cocrystals appeared to be isomorphous to the b/HLH-MLP cocrystals, I reasoned that these ends would result in G/C stacking of contiguous duplexes in the lattice which would be expected

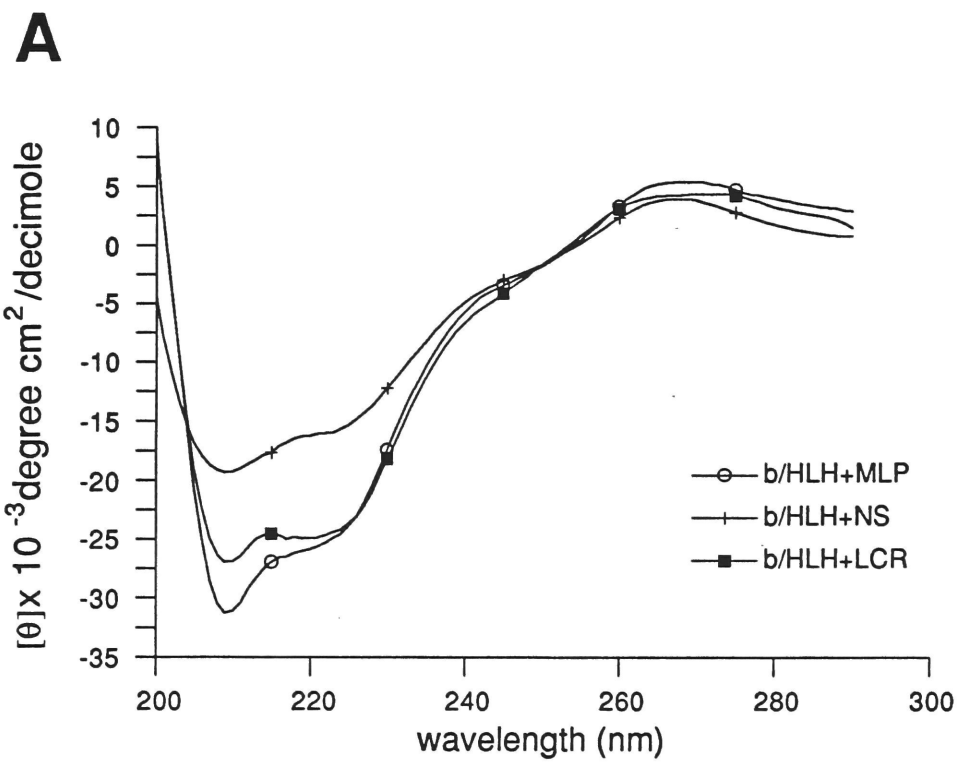
to have a free energy of staking more favorable by ~ 4 kcal/mol relative to the G/T stacking present in the b/HLH-MLP cocrystals (Saenger, 1984, p. 139).

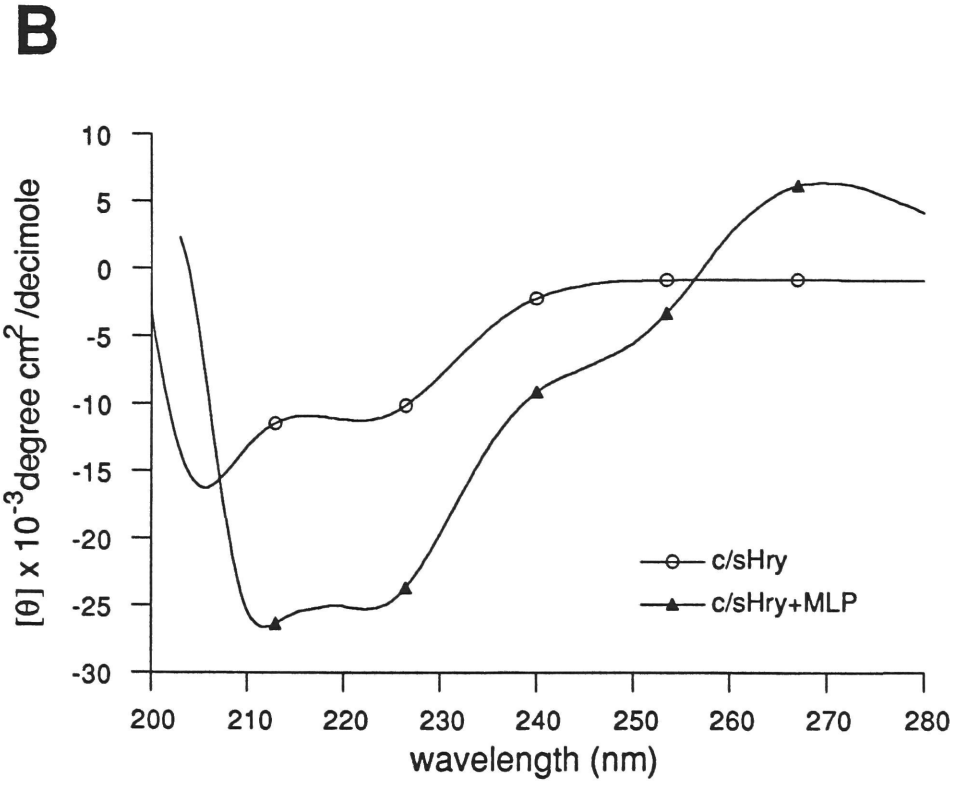
Fresh crystals grew to dimensions somewhat smaller than b/HLH-MLP crystals, and diffracted more weakly. The unit cell dimensions determined by indexing crystals held at -10 °C were a = 135.5 Å b = 54.4 Å c = 43.9 Å (the crystals appear to be isomorphous to the b/HLH-MLP cocrystals, presumably belonging to the same space group $P2_12_12_1$). An attempt at data collection was made using the Weissenberg camera (Sakabe, 1991; reviewed in Stuart and Jones, 1993) at beam-line BL6A2 of the Photon Factory (Tsukuba, Ibaraki Prefecture, Japan) with a 429.7 mm cassette, a 0.1 mm collimator, a 10° oscillation range, a coupling constant of 2 degrees/mm and an exposure time of 15 seconds/degree at 7°C. Even with this highly efficient data collection methodology, radiation induced decay limited useful data to 4Å (not shown).

Recently these cocrystals were successfully cryo-protected and flash-frozen under conditions described in Section 2.9. Data were collected from them at beam-line F-1 of CHESS. The reduced unit-cell had dimensions 135.1 x 53.5 x 43.2 $Å^3$, essentially isomorphous with b/HLH-MLP21 flash-frozen under the similar conditions. The virtually isomorphous crystallization of the LCR and MLP complexes of USF b/HLH under similar conditions suggests that the overall shape of the complex is the same. Presumably USF can bind to these two sequences by slight adjustment of its DNA interface.

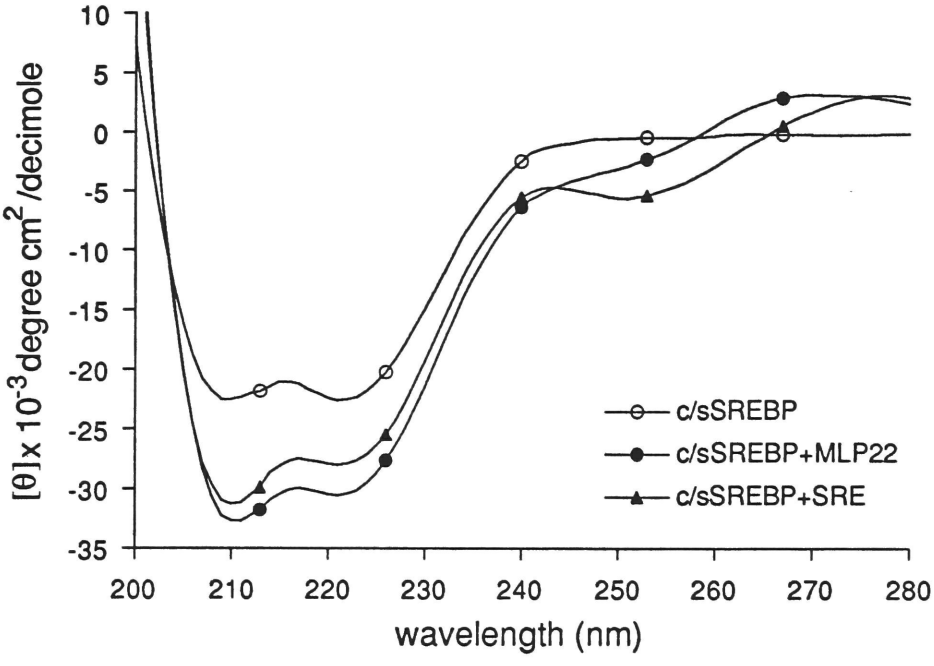
Atypical b/HLH:DNA Interactions II: Hairy Hairy is a *Drosophila melanogaster* b/HLH protein which contains a proline in the basic region (Fig. 2). It has been shown to bind DNA and to regulate fly development in a DNA-binding dependent manner (Ohsako *et al.*, 1994). Fig. 23B shows that a b/HLH construct derived from this protein (purified as described in Section 2.1) also undergoes a dramatic folding transition when challenged

Figure 23. Normalized CD spectra of USF b/HLH, c/sHRY and c/sSREBP. (**A**) Spectra of USF b/HLH in 2:1 molar mixtures with sDNA (b/HLH+MLP), with nsDNA (b/HLH+NS), and with a double-stranded 16-mer based on the sequence shown in Fig. 4B. The concentration of b/HLH was approximately 10 μM (on a monomer basis) for each measurement. (**B**) Spectra of c/sHry by itself and in a 2:1 molar mixture with sDNA (c/sHry+MLP). The concentration of c/sHry was approximately 10 μM (on a monomer basis) for each measurement. (**C**) Spectra of c/sSREBP by itself, in a 2:1 molar mixture with the MLP 22-mer shown in Fig. 4C (c/sSREBP+MLP), and in a 2:1 molar mixture with the SRE 20-mer shown in Fig. 4F. The concentration of c/sSREBP was approximately 6 μM (on a monomer basis) for each experiment.


A
10
5
0
-5
-10
-15
-20
-25
-30
-35
[θ]x 10 -3degree cm2 /decimole
200
220
240
260
280
300
wavelength (nm)
b/HLH+MLP
b/HLH+NS
b/HLH+LCR


B
10
5
0
-5
-10
-15
-20
-25
-30
[θ] x 10-3degree cm2 /decimole
200
210
220
230
240
250
260
270
280
wavelength (nm)
c/sHry
c/sHry+MLP

C


10
5
0
-5
-10
-15
-20
-25
-30
-35
[θ]x 10-3degree cm2 /decimole
200
210
220
230
240
250
260
270
280
wavelength (nm)
c/sSREBP
c/sSREBP+MLP22
c/sSREBP+SRE

with an E-box containing DNA. The protein elutes from a gel-filtration column with a elution volume similar to that of Max 22-113 and other minimal b/HLH/Z constructs (not shown). Crystallization efforts with this protein have yielded small cocrystals.

Atypical b/HLH:DNA Interactions III: the Sterol Response Element Binding Protein c/sSREBP was expressed and purified as described in Section 2.2. EMSA experiments showed it to bind both to the MLP E-box as well as to the sterol response element (SRE, data not shown), as described in the literature (Yokoyama *et al.*, 1993). CD spectroscopy corroborates (Fig. 23C) the EMSA results. The wild-type SREBP-1 b/HLH/Z construct contains a cysteine at the C-terminal end of the zipper motif which is in helical register with the leucines. Presumably because the two thiols face each other in the homodimer, the protein oxidized readily, forming a covalent dimer with reduced DNA-binding affinity, as determined by EMSA, and increased aggregation, as determined by DLS using samples fully oxidized with the copper-phenanthroline complex as a catalyst. The fully-reduced wild-type protein and the cysteine-free mutants were indistinguishable by biochemical criteria. c/sSREBP was cocrystallized with a 20 base-pair duplex oligonucleotide as described in Section 2.8. The cocrystals are in space group $C2$ with a = 154.8 Å, b = 51.8 Å, c = 46.2 Å, and β = 103.5°, and probably contains one $(c/sSREBP)_2$-DNA complex per asymmetric unit. Diffraction data have been collected from a crystal flash frozen as described in Section 2.9.

Covalently-Linked b/HLH/Z Heterodimers Covalently linked b/HLH/Z dimers were envisaged to present several advantages over b/HLH/Z proteins overexpressed in living organisms. First, they would enable production of pure heterodimers, uncontaminated with the homodimers which are unavoidable when non-covalent b/HLH/Z dimers are prepared at crystallization or NMR concentrations. This is specially true if the covalent cross-link between the protomers is asymmetric, such that only the desired

heterodimer can form. Second, total chemical synthesis allows the introduction of atom-by-atom modifications or labeling as desired, enabling single-atom level "mutagenesis" in order to study structure-function relationships or as site specific spectroscopic probes. Third, if the covalent heterodimer is prepared by the segment condensation approach, one can "mix-and-match" various segments to prepare easily an assorted library of related proteins. The total chemical synthesis of proteins (Muir and Kent, 1993), specially using the segment condensation approach, has reached the stage where synthesis was deemed to be preferable to the production of linked heterodimers in biological systems (Neuhold and Wold, 1993), which perforce incorporate a very long amino-acid linker.

Basing our design on the Max b/HLH/Z-DNA complex structure, we decided to make the covalent homo- and hetero-dimers in four segments: two b/H1 segments also containing the N-terminal half of the loops, and two H2/Z segments also containing the C-terminal half of the loops. The asymmetric covalent crosslink was introduced at the very C-terminus of the dimer, in the region of the complex which appeared to be disordered in the Max structure (Fig. 24A). As a control, a monomeric Max construct resembling Max 22-113 was synthesized by segment condensation (L. Canne, A.R.F., S.K. Burley, and S.E.B. Kent, unpublished). This protein incorporates amino acids 22 to 106 of the Max sequence with the insertion of a single thioester linkage in the loop section, between residues 53 and 54. This protein was shown to be indistinguishable from Max 22-113 in EMSA assays (not shown). CD spectra collected with this protein with and without a the E-box containing MLP 22-mer are shown in Fig. 24C.

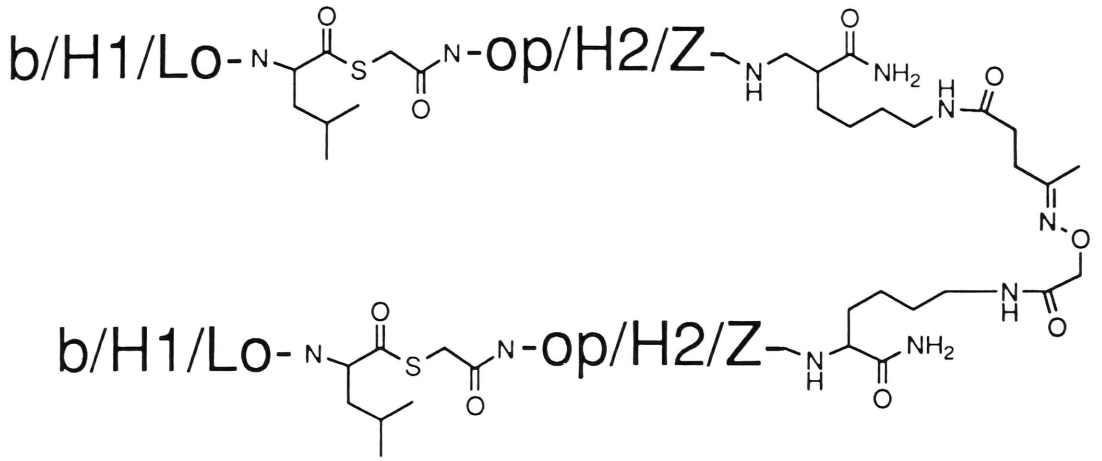
Covalently linked Max b/HLH/Z homodimers and Myc/Max b/HLH/Z heterodimers of the structure schematized in Fig. 24A were synthesized and purified in the laboratory of Stephen Kent at the Scripps Research Institute (La Jolla, California) as described elsewhere (L. Canne, A.R.F., S.K. Burley, and S.E.B. Kent, submitted). Briefly, the b/H1

segments incorporated a thio-carboxylate at their C-termini, and the H2/Z segments were bromoacetylated at their N-termini. The deprotected segments were mixed in 8 M urea at pH 4.7 to yield the thioester-ligated protomers. The C-terminal lysines of the H2/Z segments (which were protected with a different blocking group from all other lysines in the peptides, enabling their selective modification previous to deprotection of the latter, and bromoacetylation of the N-terminal amine) had been converted into the amides of either aminooxyacetic or 4-oxopentanoic acid previous to their deprotection. The keto and O-peptidylamine functionalities of the protomers were reacted under the same conditions used for thioester formation to yield the oxime crosslink at the C-termini of the protomers.

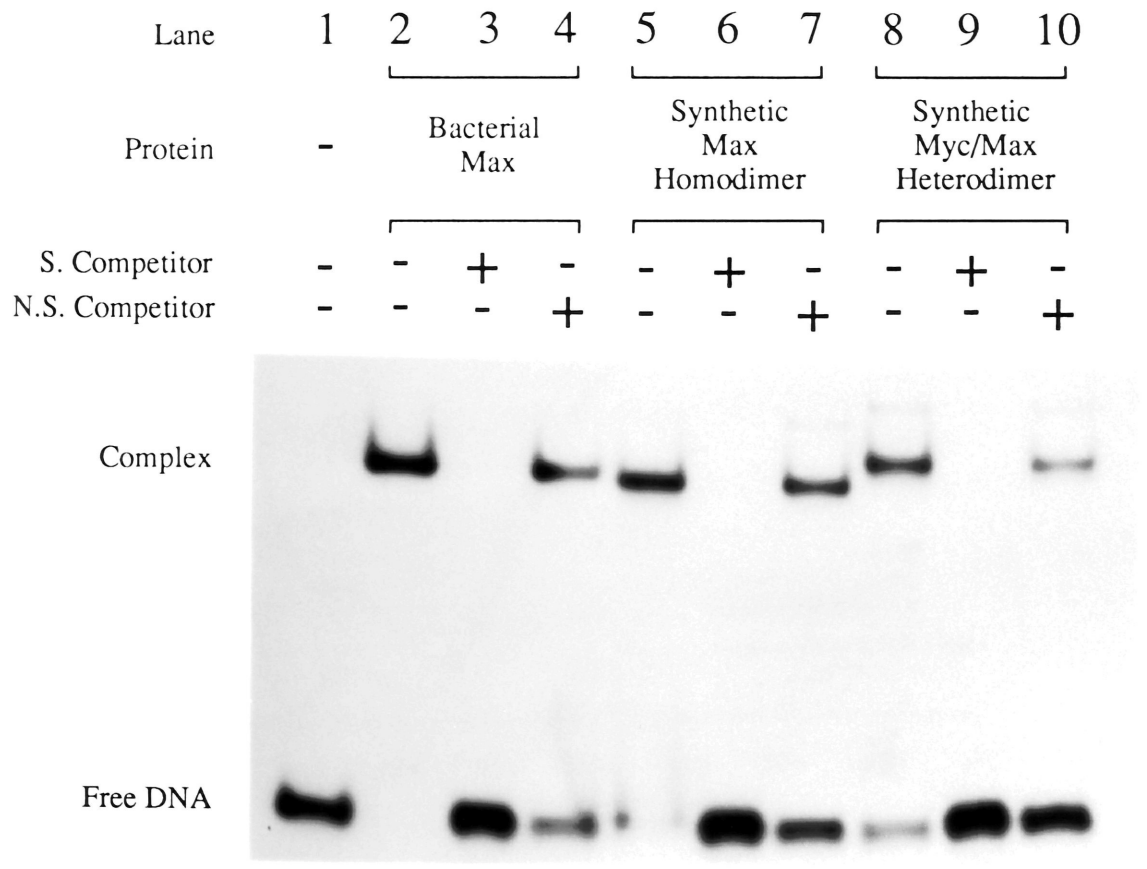
The covalently linked heterodimers are active and bind the E-box specifically as demonstrated by EMSA experiments (Fig. 24B). When the reversed-phase purified and lyophilized proteins were dissolved in buffer and immediately fractionated by gel-filtration chromatography, large aggregates eluted in the void volume. Incubation with E-box containing DNA reduced the size of the aggregates, and the protein-DNA complex had an elution volume similar to that of biologically produced Max b/HLH/Z complexed with DNA (not shown). Consistent with aggregation of the un-complexed, previously lyophilized protein, CD spectra of the covalent heterodimers resembled that of non-covalent, bacterially expressed Max 22-113, except that the helical content per residue was much higher, implying that essentially the entire protein was helical. Addition of specific DNA resulted in a molar ellipticity at 222 nm (a measure of helical content in helix+coil proteins) similar to that observed with biologically produced b/HLH/Z proteins and the synthetic monomeric Max in complex with the same DNA (Figs. 24C and 24D). Crystallization trials with these protein-DNA complexes are underway.

Figure 24. Biochemical characterization of synthetic b/HLH/Z dimers. (**A**) Schematic representation of the chemical structure of the covalent dimers. Protein segments corresponding to the canonical b/HLH/Z regions are named according to the convention of Fig. 2. The thioester linkages interrupting the loop regions and the oxime linkage at the C-terminus are shown in structural formulae. (**B**) EMSA experiment comparing bacterially expressed Max22-113 with synthetic covalent Max homodimer and synthetic covalent Myc/Max heterodimer. sDNA was used both as probe and as specific competitor; non-specific competitor was the SRE 20-mer shown in Fig. 4F. Cold competitor was present in 100-fold molar excess where indicated. The concentration of probe and protein (on a dimer basis) was 0.5 μM in each binding reaction. (**C**) Normalized CD spectra of synthetic Max protein without a covalent dimeric linkage, alone (nc-sMax) and in a 2:1 molar mixture with the MLP 22-mer of Fig. 4C (nc-sMax+MLP22). The concentration of the protein was 10 μM on a monomer basis. (**D**) Normalized CD spectra of synthetic Max covalent homodimer alone (sMax2), and in a 1:1 molar mixture with the MLP 22-mer of Fig. 4C (sMax2+MLP22). The concentration of protein was 10 μM. (**E**) Normalized CD spectra of synthetic Myc/Max covalent heterodimer alone (sMyc/Max) and in a 1:1 molar mixture with the MLP 22-mer of Fig. 4C (sMyc/Max+MLP22). The concentration of the protein was 10 μM.

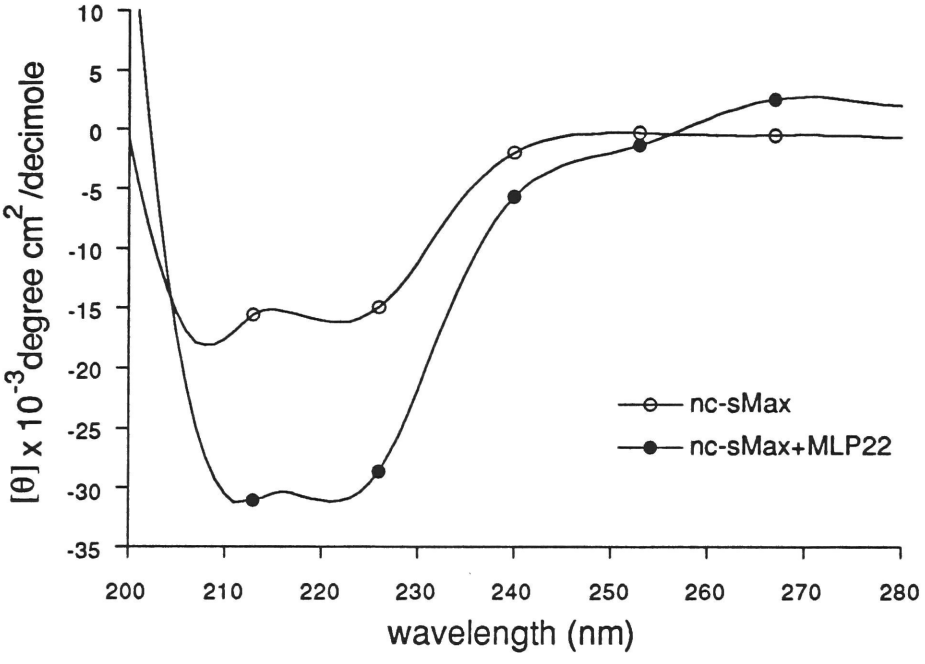
A


b/H1/Lo-
-op/H2/Z-
b/H1/Lo-
-op/H2/Z-

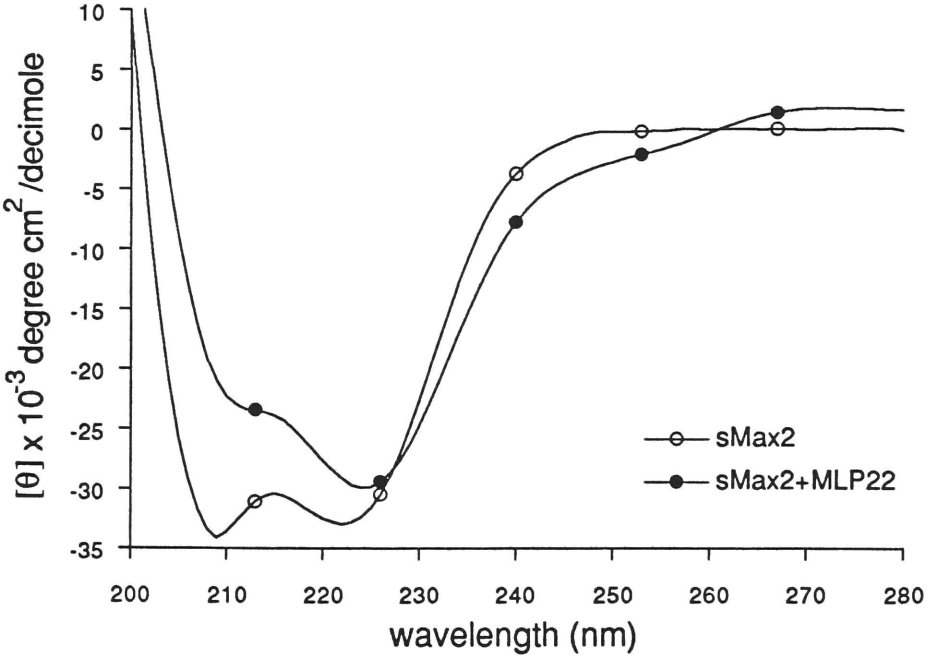
B


Lane
1 2 3 4 5 6 7 8 9 10
Protein
-
Bacterial Max
Synthetic Max Homodimer
Synthetic Myc/Max Heterodimer
S. Competitor
- - + - - + - - + -
N.S. Competitor
- - - + - - + - - +
Complex
Free DNA

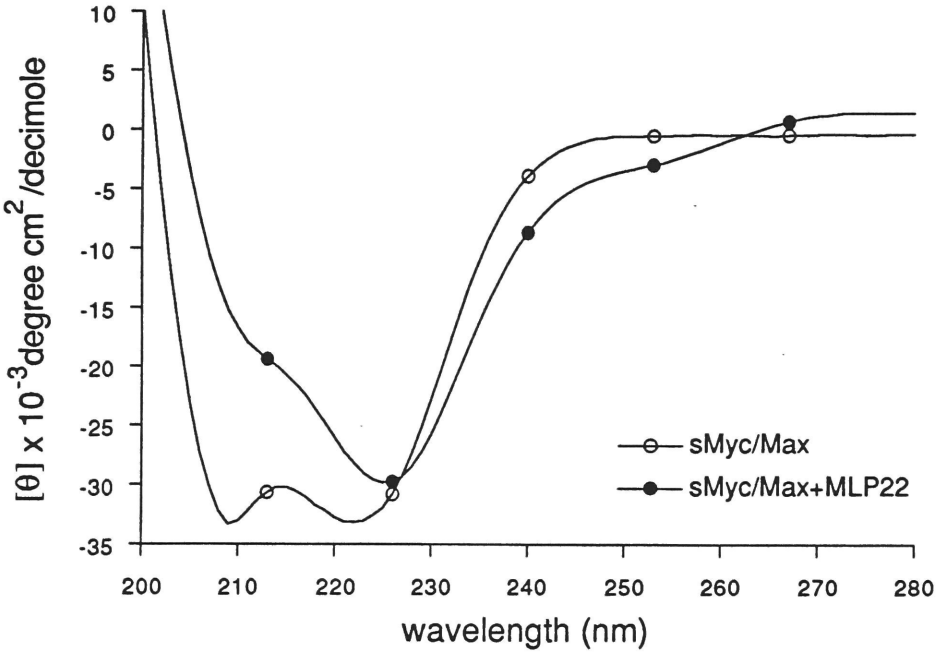
C


10
5
0
-5
-10
-15
-20
-25
-30
-35
[θ] x 10-3degree cm2/decimole
200
210
220
230
240
250
260
270
280
wavelength (nm)
nc-sMax
nc-sMax+MLP22

D


10
5
0
-5
-10
-15
-20
-25
-30
-35
[θ] x 10-3degree cm2/decimole
200
210
220
230
240
250
260
270
280
wavelength (nm)
sMax2
sMax2+MLP22

E


10
5
0
-5
-10
-15
-20
-25
-30
-35
[θ] x 10^-3 degree cm^2 /decimole
200
210
220
230
240
250
260
270
280
wavelength (nm)
sMyc/Max
sMyc/Max+MLP22

Chapter 4

Discussion

Section 4.1 discusses the novel three-dimensional structure adopted by the helix-loop-helix domain. The section starts by pointing out the limitations of the currently available crystallographic results. Then, the three-dimensional architectures of the four available helix-loop-helix structures are compared, and the structural bases of helix-loop-helix domain stability are discussed. Section 4.2 is concerned with the structural basis of sequence-specific DNA-binding by helix-loop-helix proteins. Biochemical, genetic and crystallographic results are brought to bear on the problem. Section 4.3 addresses some functional issues raised by my results. These include the possible reasons for coupling induced fit of the protein to sequence-specific DNA binding, the possible reasons for the presence of two apparently redundant dimerization interfaces in the b/HLH/Z proteins, the possible structural bases of specificity in dimerization, and the biological implications of helix-loop-helix protein tetramerization. Section 4.4 concludes this Dissertation.

4.1 A New Protein Fold: Dimerization of HLH Proteins

Limitations of the Currently Available Crystallographic Results Four structure determinations of helix-loop-helix proteins, all in complex with DNA, have been reported to date. The first report came in May of 1993: it was the structure determination of the b/HLH/Z dimer of Max bound to its target Class-B E-box (Ferré-D'Amaré *et al.*, 1993; this work). The structure determination of the USF b/HLH dimer complexed also with a class-B E-box was completed shortly thereafter (Ferré-D'Amaré *et al.*, 1994; this work). A year later, the structures of two b/HLH proteins bound to Class-A E-boxes, E47 and MyoD, were reported (Ellenberger *et al.*, 1994; Ma *et al.*, 1994). All four structure determinations were at modest resolution. The Max cocrystal structure was refined at 2.9Å; the USF structure was refined at 2.9Å, but given the completeness of the data, the effective resolution of this structure might be 3.0 Å. The E47 and MyoD were both reported at 2.8Å resolution; both refinements included a number of "water molecules" in the crystallographic model; at this resolution limit this is somewhat questionable (see *e.g.*, Karplus and Faerman, 1994), specially for the E47 structure which did not have the noise-reducing benefit of four-fold non-crystallographic symmetry averaging that MyoD had, and leads one to raise questions about the effective resolution of those structures. Overall the statements that can be made on structure-function relationships based on these crystallographic results must take into account that the precision of the atomic coordinates is at best 0.3 Å. For instance, statements about burial and qualitative proximity of side chains are valid, statements about the strength of hydrogen-bonds based on crystallographically observed distances are not.

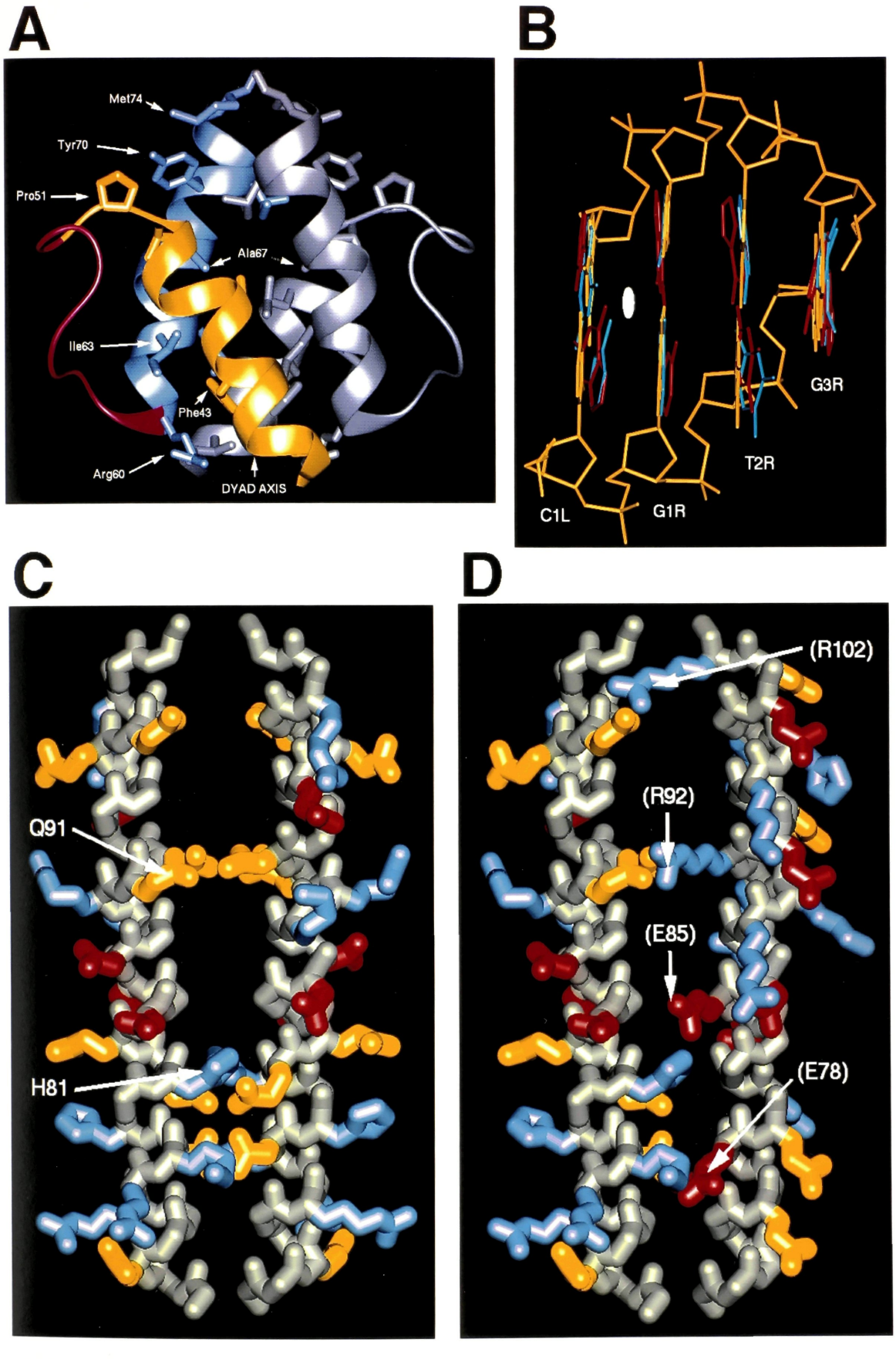
The Parallel Four-Helix Bundle Determination of the X-ray structure of a homodimer of the b/HLH/Z domain of the mammalian oncoprotein Max bound to its cognate DNA revealed that the helix-loop-helix motif dimerizes into a globular, parallel,

left-handed, four-helix bundle, which is stabilized by a well-defined hydrophobic core. As expected from the stoichiometry of DNA-protein interaction, and the deleterious effect on DNA binding of mutations in one half of the palindromic E-box (Section 3.1; Figs. 5 and 9), the dyad symmetry of the CACGTG element coincides with the dyad axis of the protein dimer. The basic region and helix 1 (H1) of the helix-loop-helix constitute a single uninterrupted α-helix, as do H2 and Z (Figure 17). Thus, the b/HLH/Z domain consists of two pairs of long α-helices, 20 and 43 residues for b/H1 and H2/Z, respectively.

There is an extensive literature on four-α-helical bundles (*e.g.*, Richmond and Richards, 1978; Weber and Salemme, 1980; Chothia *et al.*, 1981; Murzin and Finkelstein, 1988; Presnell and Cohen, 1989; Cohen and Parry, 1990; Harris *et al.*, 1994). However, in all cases the bundles considered are anti-parallel, differing only in the inter-helical connectivity or topology, and interhelical angles. Of course, parallel arrangements of four helices do occur in the interiors of other proteins, but, to my knowledge, the Max cocrystal structure represented the first report of a parallel four-helix domain as a stably folding functional unit. The packing angles between the α-helices of the HLH domain conform with the most common packing of ~ 60° (H1:H2) and ~ 19° (H2:H2) first rationalized geometrically in terms of knobs into holes interactions by Chothia *et al.* (1977) .

Although the parallel four-helix bundle was novel, it was not totally unexpected. Model building and biochemical experiments had led three groups to propose roughly correct interhelical packing geometries for helix-loop-helix proteins. Vinson and Garcia (1992) based their model building entirely on sequence alignments, and used as a strong constraint a predicted continuity of the H2 and leucine zipper helices in b/HLH/Z proteins. Halazonetis and Kandil (1992) and Davis and Halazonetis (1993) performed extensive mutagenesis on the hydrophobic core of cMyc and Max b/HLH/Z proteins, and concluded from their results that a parallel four helical bundle was the probable quaternary structure of

Figure 25. b/HLH/Z dimerization and DNA recognition. **(A)** Schematic representation of the parallel four-helix bundle formed by the Max HLH dimer. Helical regions of the protein are represented as ribbons, the loop is shown as a thin tube; side chains of conserved hydrophobic residues are shown as stick figures. One of the protomers is in gray, the other is color coded: yellow for H1, magenta for loop, and blue for H2. **(B)** Result of least-squares superposition of the E-box DNA of the Max (blue) and MyoD (red) cocrystal structures with canonical B-form DNA (yellow). The fit was performed on all C1' atoms shown using the "lsq_explicit" and "lsq_improve" commands in O (Jones *et al.*, 1991). All atoms of the canonical DNA are shown; only the bases of the crystallographically determined E-box DNA structures are shown. The sequence of the MyoD bottom strand DNA is GCTG from left to right. The approximate position of the dyad axis is indicated by the lune. See text for details. **(C)** Stick figure representation of the leucine zipper region of Max. Acidic residues are shown in red, basic in blue, polar in yellow; hydrophobic residues are not shown. **(D)** Same as (C), but with the side chains of the molecule on the right hand side substituted for the residues situated at equivalent positions in the cMyc sequence. The position of backbone atoms was left unchanged. The rotamers shown are the most common ones (Jones *et al.*, 1991). No attempt was made to optimize side chain interactions. Myc side chains are numbered using the Max scheme (Fig. 2).


A
Met74
Tyr70
Pro51
Ala67
Ile63
Phe43
Arg60
DYAD AXIS
B
G3R
T2R
C1L
G1R
C
Q91
H81
D
(R102)
(R92)
(E85)
(E78)

b/HLH/Z dimers. Anthony-Cahill *et al.* (1992) determined the parallel orientation of the H2 segments of b/HLH peptides by introducing a nitroxide spin label at their C-termini and confirming their proximity (separation of the labels was judged to be less than 20Å) by EPR spectroscopy. All three groups, however, made incorrect right-handed, instead of the correct left-handed, connections between H1 and H2. In addition, Vinson and Garcia placed the N-termini of the H2 regions inside the major groove of DNA, making contacts with the base edges, while Halazonetis and Kandil based their model on the *Eco*RI protein structure, assigning an extended-chain conformation to the basic region polypeptide backbone. Neither the protein-protein nor protein-DNA interfaces were correctly predicted in any detail by these three groups of workers.

Two groups proposed the wrong anti-parallel four helix bundle as the likely structure of helix-loop-helix proteins. Gibson *et al.* (1993) based their modeling on sequence alignments and the structure of the Rop RNA binding protein, whose antiparallel four helix-bundle structure has been determined crystallographically. Starovasnik *et al.* (1992) based their model on an incomplete NMR study of a disulfide crosslinked MyoD b/HLH peptide. This oxidized form of MyoD is inactive in DNA binding (Starovasnik *et al.*, 1992; Ma *et al.*, 1994); I propose that the NMR data probably correspond to an alternative packing of the HLH domain which is not compatible with DNA binding

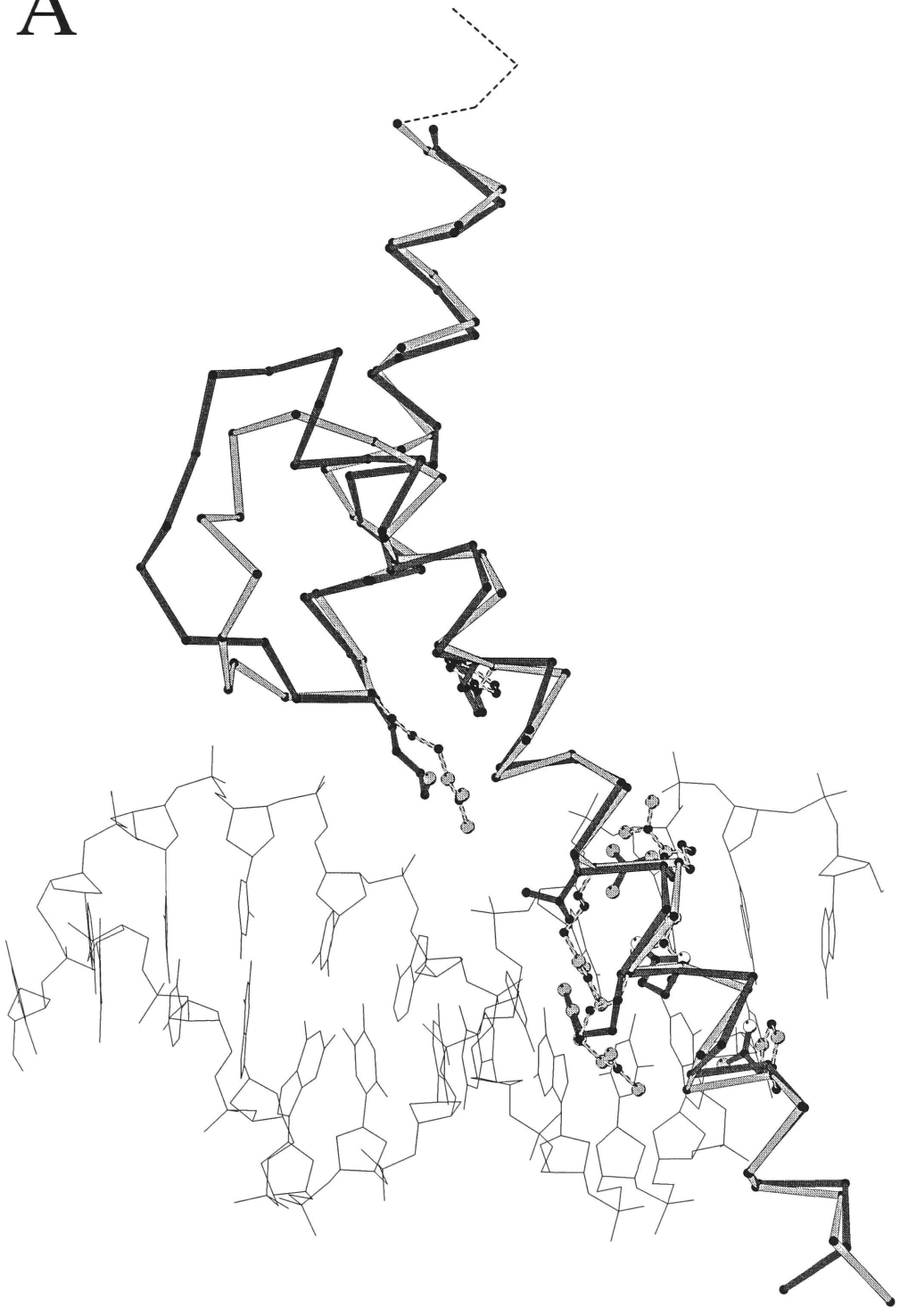
The orientation of the H2/Z α-helices is such that their N-termini lie towards the DNA. α-helices have partial positive charges in their N-terminus because of the alignment of the backbone carbonyls towards the C-terminus (reviewed in Hol, 1985). This may result in a favorable energetic contribution of the apposition of the N-terminus of the H2/Z helix on the negatively charged phosphate backbone of the DNA. A substantial body of literature has grown around the "helix macrodipole", the idea that the partial positive charge mentioned above is additive and that in the case of long α-helices, a substantial electric

dipole is generated. This idea was invoked in the past to explain why only anti-parallel four helix bundles had been observed (Hol, 1985). Recent experiments (He and Quiocho, 1993) and calculations (Åqvist *et al.*, 1991) have laid to rest the notion of the macrodipole; fluctuations in the dielectric properties of the environment of the helix result in the N-terminal partial positive charge being the same regardless of number of α-helical turns.

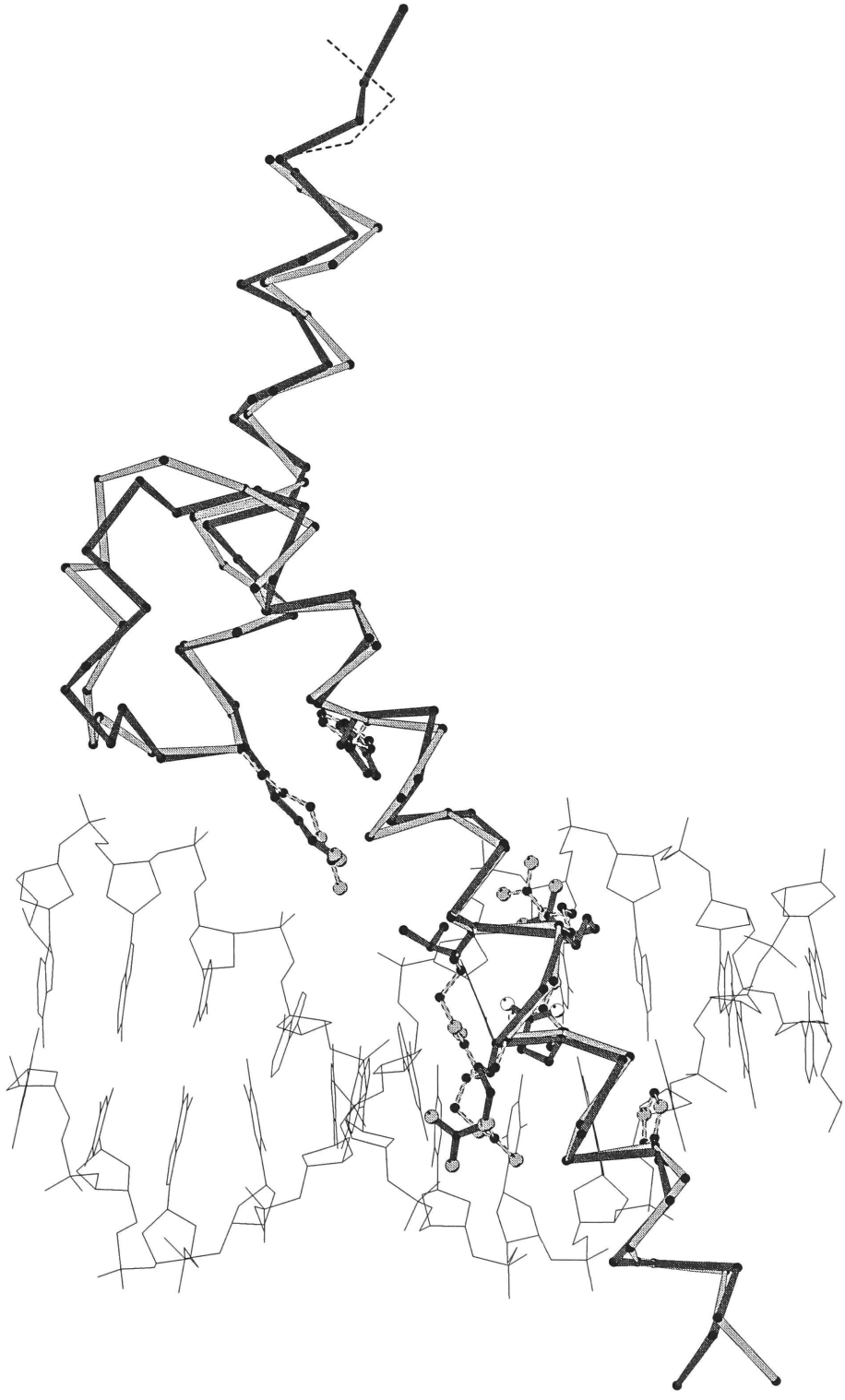
Comparison of b/HLH/Z and b/HLH Structures As described in some detail in Section 3.3, the hydrophobic core of the globular four-helix bundle is formed by residues on the non-polar faces of the four amphipathic helices, which are highly conserved between HLH, b/HLH and b/HLH/Z proteins (Figs. 2, 20A, 20B, 20C, 20E, 25A). This striking sequence conservation and the results of model building led me to predict that b/HLH proteins would have the same tertiary and quaternary structures as Max (Ferré-D'Amaré *et al.*, 1994). Subsequent structure determinations of the DNA-binding domains of the b/HLH domains of E47 (Ellenberger *et al.*, 1994), and MyoD (Ma *et al.*, 1994), as well as the structure determination of the truncated b/HLH form of the DNA binding domain of USF (Ferré-D'Amaré *et al.*, 1994; this work) confirmed the prediction. As can be seen in Figs. 26A and 26B, the conformations of the Max, E47 and MyoD four-helix bundles are virtually identical. Max and MyoD were aligned using the α-carbons of 29 residues (D23 through P51) from b/H1 and 26 residues (G55 through T80) of loop/H2 giving a rmsd of 1.2 Å. Max and E47 were aligned in the same way, using the same 29 residues from b/H1 and 21 residues (K57 through K77) from loop/H2 with a rmsd of 1.0 Å. The solvent accessible surface area buried upon dimerization of the HLH domain of the MyoD structure differs by less than 2 % from that buried by dimerization of the same domain of Max. All the conserved hydrophobic residues make equivalent contacts in these proteins. As expected, the packing of the loops differ somewhat, but the backbone geometries are very similar in the last two amino acids of the loop region. This is accounted for at least in part by the contacts made with the DNA by the last residue in the loop region.

Figure 26. Three-dimensional alignment of the Max b/HLH domain with E47 and MyoD. Both figures were prepared with MOLSCRIPT (Kraulis, 1991). (**A**) Least-squares alignment of Max and E47 b/HLH regions. Coordinates for the E47-DNA complex were kindly provided by T. Ellenberger (Harvard Medical School). Alignments were performed using all α-carbons shown with the same methodology described in Fig. 25B. The α-carbons of E47 are connected with dark gray tubes, those of Max with light gray; some key E47 side chains are in dark gray; the corresponding Max side chains are shown in light gray with dotted contours. The DNA from the E47 model is represented with thin lines. The DNA shown corresponds to a full asymmetric unit of the structure as determined by Ellenberger *et al.*, 1994. The E47 protomer shown corresponds to the GTG half-site. (**B**) Same as (A) but with MyoD instead of E47. Coordinates of the MyoD complex were kindly provided by C. Pabo (Massachusetts Institute of Technology). The DNA shown is that of the MyoD structure. The crystallographic model for that structure (Ma *at al.*, 1994) contains two 14-mer DNAs in each asymmetric unit.

A



B



The H1 region of the E47 structure extends one helical turn beyond the last helical residues of the H1 regions of the other three structures. This is possibly a consequence of this protein both not having a proline at the end of H1, and not having a residue equivalent to Tyrosine 70 in the outer surface of H2; this last residue packs against the top of H1 in the other three structures (see Figs. 20A, 20C, 22D, 25A). It is unknown if there are any energetic consequences of having this extra turn. Clearly it does not affect the structure of the remainder of the HLH domain (Fig. 26A). Other proteins which lack a proline at the end of H1 (such as Hairy, Fig. 2) may also have this extension of the helix. The different interhelical packing angle observed in the USF b/HLH structure (Section 3.4) will be discussed in Section 4.3. Based on of the high degree of structural conservation evident in the helix-loop-helix structures, I will use the structure of Max which I described in some detail in Section 3.4 for interpreting mutagenesis results obtained from diverse b/HLH and b/HLH/Z proteins.

Concordance with Mutagenesis Results Because of their high degree of phylogenetic conservation, mutagenesis results showing that gross alteration (Davis *et al.*, 1990; Davis and Halazonetis, 1993) of the hydrophobic residues forming the core of the four-helix bundle abolished dimerization, and consequently DNA binding, are not very illuminating. Point mutation results are more interesting, and all highlight the exquisitely tuned close packing of the core of the HLH domain. Voronova and Baltimore (1990) made a point substitution of the phylogenetically conserved basic first residue of H2 (lysine in E47) to alanine; this resulted in loss of dimerization and DNA binding. In light of the available structural information, it is clear why this is so detrimental. The lysine they mutated corresponds to Arginine 60 in Max; this residue stabilizes (1) the H1:H2 interaction by packing of its methylene groups with the phylogenetically conserved bulky hydrophobic residue in the apposing face of H1 (Phenylalanine 43 in Max) as can be seen in Figs. 13A, 20A, 20C, 25A; and (2) makes a multidentate contact with DNA, anchoring

the four-helix bundle on the phosphodiester backbone with the correct registration for productive binding (Figs. 20E, 21, 26A, 26B.) Equivalent protein-protein and protein-DNA contacts of this residue are present in the USF, E47, and MyoD structures.

Shirakata *et al.* (1993) mutagenized the bulky hydrophobic corresponding to Phenylalanine 43 of Max in MyoD, and assayed the mutant proteins' ability to dimerize and bind DNA. They found that the apparent affinities for DNA decreased in the order Phe (wt) > Tyr > Leu > Ile > Val > Ala; the protein with alanine was inactive. It is clear from the crystal structures that the mutations were introducing subtle or gross distortions in the packing of the four-helical bundle core, with the bulkier residues being better tolerated. In fact, wild-type USF has an isoleucine in this position. A mutation of the corresponding phenylalanine in E47 to a glutamic acid predictably abolished dimerization (Voronova and Baltimore, 1990). Introduction of a kink in the H2 α-helix by substitution of the residue corresponding to Leucine 64 of Max in the myogenic b/HLH factor Myf-5 (a leucine too) with a proline completely abolished dimerization and DNA binding (Winter *et al.*, 1992). This residue is located in the hydrophobic core, just opposite Phenylalanine 43 (Fig. 20C, 25A). Introduction of a kink in H1 by substituting the residue corresponding to Serine 45 of Max in Myf-5 (a threonine) with a proline also completely abolished dimerization. A double mutant of E47 in which the residues corresponding to Isoleucine 71 in Max (isoleucine in E47) and Methionine 74 in Max (leucine in E47) were changed into aspartate and lysine, respectively, resulted in a protein incapable of dimerizing. These residues are at the H2:H2 interface as can be seen in Fig. 20C. Naïvely, one might imagine that a pair of salt bridges would form, allowing dimerization to occur, but the experimental result is that such a pair of ion-paired residues is too bulky for stable dimerization.

The less stringently conserved residues on the outer edge of the hydrophobic core are more tolerant of mutations. The isoleucine which in MyoD and Myf-5 occupies the position

corresponding to Threonine 68 in Max (Fig. 20C) sits at the upper end of the H1:H2 junction, away from the loop, on the border between solvent and hydrophobic interior. This residue tolerated changes into alanine or phenylalanine in MyoD (Shirakata *et al.*, 1993), and into valine in Myf-5 (Winter *et al.*, 1992).

Residues in the loop can be changed quite liberally in number and composition without compromising dimerization or DNA-binding. The shortest length compatible with function is probably five residues. Starovasnik *et al.* (1992) found that MyoD b/HLH constructs stopped binding to DNA when the loop was shortened from six to four residues. The shortest naturally occurring loop is five residues long (CBF1, Cai and Davis, 1990). The distance between the last α-carbon of H1 and the first α-carbon of H2 is fixed at 15 Å by the Max structure. Given the Cα-Cα distance of 3.8 Å between contiguous residues in an extended polypeptide chain, five residues appears to be just enough to comfortably bridge the gap. Lengths of more than thirty occur in some *Drosophila* b/HLH proteins of the Achaete-Scute subfamily (see references in Benezra *et al.*, 1990). The hydrophobic residues present in the loop pack against the outer face of the HLH core, as can be seen in Fig. 20E, but the results of mutagenesis imply that this packing is incidental, not required for stability of the domain.

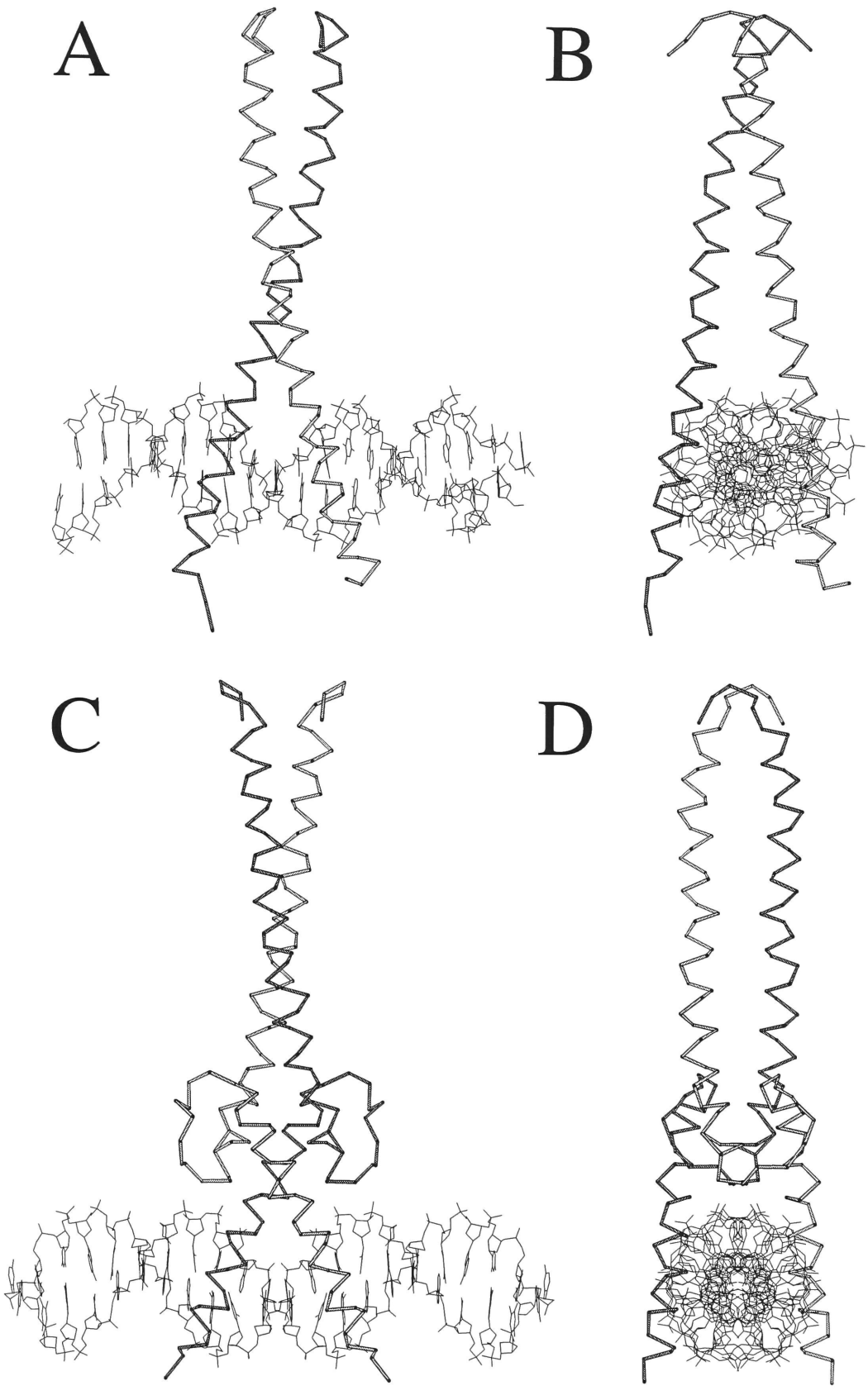
The four cocrystal structures also provide some insights into the mechanism of action of HLH proteins (Id and emc), which lack the basic region, do not bind DNA and function as negative regulators of b/HLH proteins by forming heterodimers that are defective in DNA binding (Benezra *et al.*, 1990; Ellis *et al.*, 1990). The helix-loop-helix regions of these proteins are very similar to those of both the b/HLH and b/HLH/Z proteins (Fig. 2), and heterodimerization of HLH with b/HLH proteins probably relies of forming the same four-helix bundle seen in the cocrystals structures of the b/HLH and b/HLH/Z homodimers. In

isolation, the helix-loop-helix motif of Id forms a stable homotetramer (Fairman *et al.*, 1993). The physiological relevance of this oligomerization state has not been established.

Additional biological illustrations of the dominant negative properties of DNA-binding defective b/HLH/Z proteins are afforded by a number of recently-characterized alleles of the mouse *microphthalmia* locus. For instance, the Mi^{or} allele, which has a single point mutation transforming Arginine 216 (equivalent to Arginine 35 in Max) into a lysine results in a protein which is defective in DNA-binding, and is semidominant. The mi^{ce} allele, which introduces a stop codon after the H2 region and results in a protein missing the heptad repeat, is, as would be expected from a mutation resulting in a protein deficient in dimerization, recessive (Steingrímsson *et al.*, 1994).

Comparison with the GCN4 b/Z Structure The three dimensional structures of the Max b/HLH/Z-DNA complex (Ferré-D'Amaré *et al.*, 1993; this work) and the GCN4 b/Z-DNA complex (Ellenberger *et al.*, 1992) are illustrated side-by-side on Figs. 27A - 27D. The basic region-leucine zipper (b/Z) domain is composed of a single pair of extended α-helices. Although at first sight the architecture of the b/HLH/Z and b/HLH proteins might appear to resemble the purely coiled-coil b/Z proteins, the presence of a well-defined globular core distinguishes the helix-loop-helix proteins. In particular, the four-helix bundle appears to confer an orientation closer to parallel with the DNA helical axis on the basic regions of these proteins, which explains why b/HLH/Z (and b/HLH) proteins almost invariably recognize inverted repeats (palindromes) without spacing between the repeated elements; in contrast, the b/Z proteins, whose two extended α-helices splay towards the N-terminus, usually bind inverted repeats with a variable spacing of one or two base pairs (reviewed in König and Richmond, 1993). The features that these families of proteins do share are their α-helical secondary structure, and the presence of basic region ("recognition") α-helices which are stabilized by coming into contact with DNA; the

Figure 27. Comparison of the Max b/HLH/Z-DNA complex with the GCN4 b/Z-DNA complex. MOLSCRIPT (Kraulis, 1991) figures show α-carbons connected with tubes; DNA is represented with thin lines. **(A)** The GCN4 b/Z-DNA complex (Ellenberger *et al.*, 1992) viewed perpendicular to the helical axis of DNA. Atomic coordinates kindly provided by T. Ellenberger (Harvard Medical School). **(B)** View of GCN4 rotated 90° relative to (A). **(C)** Max b/HLH/Z-DNA complex viewed at an orientation equivalent to that of (A). **(D)** Same as (C) but rotated 90°.


A
B
C
D

recognition helix is not sandwiched between a globular domain and the DNA as observed in most protein domains which use an α-helix for recognition of major groove nucleotide base edges (*e.g.* the HTH proteins, Fig. 1A, or the nuclear receptor DNA-binding domains, Fig. 1C).

Comparison of Figs. 27A and 27C demonstrates the different angle made by the basic regions of these proteins with respect to the DNA helical axis. The b/HLH/Z recognition helix is about 60° away from the DNA helical axis, while the corresponding b/Z helix makes an angle of approximately 75°. The other important consequence of the presence of a globular core in the helix-loop-helix proteins is that since the C-terminal helices (H2/Z) in these proteins do not enter the major groove, but, rather, are apposed onto the DNA backbone, these α-helices do not have to splay towards their N-termini in the way the b/Z helices do. This is made specially clear by comparing Figs. 27B and 27D. As pointed out initially by Crick (1953b) the non-integral number of residues per helical turn that characterizes the α-helix (Pauling *et al.*, 1951) implies that a pair of helices interacting through a periodically repeated set of hydrophobic residues on one of their faces will have to wrap around each other in order to maintain contact; this results in the characteristic superhelical pitch of coiled coils (reviewed in Seo and Cohen 1993). Figs. 27A-27B show that the b/Z coiled coil helices wrap around each other in their C-terminal ends, as necessitated by their having to accommodate the DNA duplex in their N-terminus. In contrast, the region of closest contact between the H2/Z helices of Max is in the H2 region which b/HLH and b/HLH/Z proteins share. The more open nature of the leucine zipper region of b/HLH/Z proteins compared to the homonymous region of b/Z proteins is reflected in cMyc and Max tolerating a three amino acid insertion in the H2-Z boundary (Davis and Halazonetis, 1993), in the somewhat non-standard presence of a number of polar residues in the Max leucine zipper, and in the fact detectable in the alignment of b/HLH/Z proteins in Fig. 2 that the phasing of leucines (which should occupy position "d"

of the canonical leucine zipper) appear in either the canonical position "d" as throughout the Max, cMyc, and Mad proteins' zippers or change from position "d" to the other interfacial position, "a" in the USF zipper. This shift in phase is exactly what the three amino acid insertion mentioned above would have brought about in Max and cMyc.

4.2 DNA Recognition by b/HLH Proteins

Comparison with Mutagenesis Results Inspection of the Max cocrystal structure and posterior survey of the four currently available cocrystal structures (Figs. 17, 22, 26A, 26B) indicated that three portions of the b/HLH segment make DNA contacts: the basic region, the loop, and the first residue of H2. As mentioned above, results of site-directed mutagenesis studies of the b/HLH proteins E47 (Voronova and Baltimore, 1990), MyoD (Davis *et al.*, 1990; Starovasnik *et al.*, 1992; Shirakata *et al.*, 1993), and Myf-5 (Winter *et al.*, 1992), and the b/HLH/Z heterodimer Myc/Max (Davis and Halazonetis, 1993), suggested that loop sequence and amino acid composition have little if any effect on DNA binding. The E47 cocrystal structure demonstrates that its 7 residue loop does not make any contacts with DNA. In contrast, three of the four cocrystal structures show loop-DNA interactions: a loop to distal (across the minor groove) contact for the Max 8 residue loop (Fig. 20E), two distal phosphate contacts and one proximal minor groove sugar contact for the USF 12 residue loop (Fig. 22D), a single proximal phosphate contact for the MyoD 8 residue loop (Fig. 26B). The very last residue of the loop consensus region of all four proteins makes either a backbone or side-chain phosphate contact. This is exemplified in the Max structure by the contact made by Serine 59 (Fig. 20E). It is possible that the very long loop regions present in some b/HLH and b/HLH/Z proteins will fold into small domains of their own, and participate in protein-DNA or protein-protein interactions.

There are two invariant residues in the basic regions of all b/HLH and b/HLH/Z proteins capable of binding E-boxes, Glutamate 32 and Arginine 35 (in the Max numbering scheme). The four cocrystal structures implicate these two amino acids in recognition of the outer two base-pairs of the E-box, CAnnTG. In each half site of the Max-DNA complex, the glutamate side chain participates in hydrogen bonds with the outer cytosine (N4) of the palindrome and the middle adenine (N6) and thymine (O4). Arginine 35 positions the glutamate side chain relative to the DNA backbone and neutralizes its negative charge, forming salt-bridges between the Oε and Nε and Nη and the phosphodiester backbone, respectively. Identical interactions are present in the USF, MyoD and E47 structures, confirming that these two residues play a critical role in recognizing the outer two base pairs of the E-box. In addition, and as can be readily appreciated from Fig. 13B, the methylenes of the Glutamate 32 side-chain desolvate and make van der Waals contact with the methyl group of the thymine residue of the E-box. To my knowledge, no one has directly tested the energetic contribution of this contact by measuring the affinity of a helix-loop-helix protein for a DNA containing a deoxyuridine in this position.

The importance of the multidentate interactions made by these two residues is underscored by the results of site-directed mutagenesis and *in vivo* genetic studies. Fisher *et al.* (1993) and Feldman *et al.* (1993) found that substitution of Glutamate 32 for alanine results in loss of DNA binding, while Fisher and Goding (1992) found that substitution of the glutamate by leucine, aspartate or asparagine abolishes DNA binding (substitution to glutamine was tolerated by the Pho4 yeast b/HLH protein). Davis *et al.* (1990) found that substitution of Glutamate 32 for aspartate completely abolished DNA binding and transactivation by MyoD. Alanine and leucine cannot hydrogen bond, and aspartate or asparagine are too short to reach into the deep major groove. Voronova and Baltimore (1990), and Fisher *et al.* (1993) found that substitution of Arginine 35 by lysine abolishes DNA binding, confirming the importance of the bidentate interactions made by Arginine 35. An

interesting biological correlate of this effect on DNA binding was found during characterization of the Mi^{or} semidominant allele of the mouse microphthalmia locus, which results in small or absent eyes and osteoporosis in the homozygote and is caused by an arginine to lysine point mutation in this position of the b/HLH/Z protein encoded by Mi^{or} (Steingrímsson *et al.*, 1994).

The Max cocrystal structure reveals a hydrogen bond between N7 of the outermost guanine of the E-box and Histidine 28 (Figure 4), and equivalent interactions occur in USF (histidine) and E47 (asparagine). Although MyoD has an alanine in this position, an arginine one helical turn towards the N-terminus makes an analogous base contact. Hydrogen bonds with N7 alone specify a purine base, but together in this context Glutamate 32 and Histidine 28 recognize a G:C base pair. In Max and USF, the imidazole ring also makes hydrophobic contacts with the 5-methyl group of the central thymine of the E-box, as suggested by the photocrosslinking of the histidine to a 5-bromo-uracil in a nonphysiologic Myc-DNA complex (Dong *et al.*, 1994), and cMyc/Max b/HLH/Z complexes (R. Ebright, personal communication.) The methyl group of the thymine appears to be almost completely desolvated by the contacts with Glutamate 32 and Histidine 28 (Fig. 18I). In MyoD, a nearby threonine residue (position 29 in the Max numbering system) makes similar hydrophobic interactions with this methyl group. Max and USF both have an asparagine at position 29 that makes hydrogen bonds with the phosphate backbone (Fig. 20F).

Class-B b/HLH and b/HLH/Z proteins recognize the E-box caCGtg. In both the Max and USF structures (Figs. 21 and 22E), Arginine 36 (Arginine 212 in USF) donates a hydrogen bond to N7 of the guanine closest to the dyad axis of the complex. Max homodimers and Myc/Max heterodimers will bind to either caCGtg or caTGtg (Blackwell *et al.*, 1993), which is consistent with the lone Arginine 36-purine N7 contacts seen in the

Max and USF structures. However, this structural feature cannot be the sole determinant specifying the central two base-pairs of the E-box, because Max and other Class-B proteins do not bind efficiently to caTAtg E-boxes (Blackwell *et al.*, 1993; Feldman *et al.*, 1993). Class-A proteins recognize caGCtg or caGGTg E-boxes, and amino acid sequence comparisons and site-directed mutagenesis suggest that they do so because there is a hydrophobic residue at position 36 instead of an arginine (Dang *et al.*, 1992). The cocrystal structures of the Class-A proteins E47 and MyoD show that the corresponding valine and leucine are far from the central base-pair and do not interact directly with DNA. E47 was cocrystallized bound to the asymmetric site CAGGTG. In the GTG half-site, the arginine corresponding to Arginine 33 in Max (where it makes phosphate contact, Fig. 20F) moves into the major groove space left vacant by replacing the Max Arginine 36 with a valine in E47, and makes a hydrogen bond with the N7 of guanine (Fig. 26A). The CAG half-site does not, of course, show such an interaction, having a large unfilled space in the major groove as does MyoD (cocrystallized with a CAGCTG site.) Thus, neither Class-A nor Class-B DNA-binding specificity can be explained entirely in terms of a readout mechanism employing direct amino acid-base contacts in the major groove.

A perplexing feature of many of the basic regions of b/HLH/Z proteins is the presence of a solvent exposed hydrophobic residue in the position occupied by Leucine 31 in the basic region of Max (Fig. 20F; compare with the alignment in Fig. 2.) What the role of this amino acid is remains unknown, but at least in the case of the microphthalmia locus product, substitution of the wild-type isoleucine for asparagine results in a phenotype in the organism: Mi^{wh} homozygous mice have white coat, small eyes, and have inner ear and mast cell deficiencies; (Steingrímsson *et al.*, 1994). Since the hydrophobic residue is partially exposed in the bound complex, it may be used to interact with other components of the transcriptional machinery. Alternatively, its presence may be biologically important

because it *lowers* the affinity of the protein for DNA; I will return to this idea in Section 4.3.

DNA Conformation and Specific Binding As discussed in Section 3.4, the structure of the DNA present in the Max cocrystal does not deviate significantly from that of canonical B-form DNA. The same is true of the DNA in the other three cocrystal structures. The MyoD E-box DNA segment can be superimposed onto the Max E-box with an rmsd of 0.55 Å for the C1' atoms. In turn, the Max E-box DNA can be superimposed with canonical B-form DNA with an rmsd of 0.56 Å for the same atoms. The superimposed DNAs are shown in Fig. 25D. It can be appreciated that the only systematic deviations present are the high roll, propeller twist and buckling of the A:T base pair, and a slight compression of the major groove, as evidenced by the roll of the central G:C and C:G base pairs towards the dyad axis. This essentially straight B-form conformation of the DNA is in marked disagreement with earlier biochemical experiments purporting to show DNA bending by Max and related proteins ranging from 50 to 80° (Wechsler and Dang, 1992; Fisher *et al.*, 1992). The presence of straight B-form DNA in four cocrystal structures, each with a different kind of crystal packing, implies that the bending observed in phasing studies must be a higher order effect, possibly a result of tetramerization.

How then do b/HLH and b/HLH/Z proteins recognize their cognate DNA? The proteins may be employing sequence-dependent DNA deformability to recognize their targets. The high density of basic region-DNA backbone contacts (Ferré-D'Amaré *et al.*, 1994), and the characteristic deviation of the E-box from canonical B-form DNA (the buckling of the A:T base pairs, a narrowing of the major groove and a concomitant widening of the minor groove) seen in all four cocrystal structures (Ma *et al.*, 1994), suggests that this could indeed be the case. The precise role of water molecules in E-box recognition is not yet established, because of the moderate resolution limits (2.8-2.9Å) of the current structures.

Clearly, as underlined in discussing other protein-DNA complexes in the Introduction, efficient desolvation of both the nucleotide base edges and the phosphodiester backbone takes place upon association of helix-loop-helix proteins with DNA; put another way, the E-box DNAs and the basic regions have a high degree of surface shape complementarity. The detailed breakdown of the enthalpic contributions of the various interactions detected in the cocrystal structure remains to be investigated.

4.3 Biological Function of HLH Proteins

Why a Folding Transition? Circular dichroism spectroscopic investigations of DNA binding by the isolated b/HLH domain of MyoD (Anthony-Cahill *et al.*, 1992), the b/HLH/Z domain of TFEB (Fisher *et al.*, 1993), full-length USF as well as its isolated b/HLH/Z and truncated b/HLH DNA-binding domains (Ferré-D'Amaré *et al.*, 1994), and a variety of other b/HLH/Z and b/HLH constructs (reported herein), demonstrated that these proteins undergo a dramatic random coil to α-helix folding transition upon association with their cognate DNA. The magnitude of the folding transition (about twenty residues per protein monomer) and the fact that the only α-helical region of the complexed proteins in contact with the E-box bases of the DNA is the basic region, imply that the basic regions are unfolded in the absence of E-box DNA and fold during DNA recognition. A similar coil-helix transition has been observed in the basic regions of b/Z proteins (Weiss *et al.*, 1990). Induced fit appears to be a hallmark of specific DNA binding by some proteins (Spolar and Record, 1994); the b/HLH, b/HLH/Z, and b/Z proteins provide some of the most extreme examples of this phenomenon.

On "architectural" grounds, it is clear why the basic regions are disordered in the absence of specific DNA. These polypeptide segments with a large net positive charge are not stabilized by tertiary or quaternary interactions with other parts of the protein, so no

hydrophobic interactions can pay for the coulombic price of holding a group of positive residues together in a helix, or the entropic cost of collapsing an ensemble of random polypeptide conformations into a tightly restrained ensemble of helical conformations. Other DNA-binding domains, such as the HTH and the nuclear receptor subfamily zinc finger domains, have equally basic "recognition helices" (see alignments of different DNA-binding protein families in Pabo and Sauer, 1992) but the globular cores of those proteins stabilize the helical structure of those basic polypeptide segments in the absence of DNA. In order to recognize specific DNA sequences by inserting isolated α-helices into the major groove, coupling folding to binding is probably inevitable.

Basic residues have the lowest helical propensities of all amino acids in the context of the Zimm-Bragg statistical mechanical formalism of coil-helix transition (Zimm and Bragg, 1959). The helical propensities of amino acid residues have recently been shown to be accountable purely on entropic grounds: the more restrictions the helical conformation of the backbone introduces on the alternate conformations a side-chain may adopt relative to the conformations it assumes when the backbone is in an extended conformation, the lower the helical propensity (Creamer and Rose, 1992; Pickett and Sternberg, 1993). The α-helical propensities of the amino acid side chains were calculated by these authors by surveying the number of alternate conformations seen to be adopted by side chains in helical and non-helical regions of well-refined, high-resolution crystal structures of proteins, and then using statistical mechanical formalism (analogous to that used to compute the information content of DNA binding sites, Chapter 1) to compute the entropic price; the results were in excellent agreement with experimentally measured propensities. Not unexpectedly, the long, basic side chains suffer a considerable reduction in accessible number of conformations upon helix formation, and are computed to have the lowest helical propensities. The entropic cost of forming the basic region helices of b/HLH, b/HLH/Z and b/Z domains is then even higher than that for an "average" composition

polypeptide, and the favorable enthalpic and entropic gains resulting from charge neutralization, the formation of polar contacts, and burial of hydrated non-polar surface area achieved by the formation of the specific basic region-E box complex, must all be expended to fold this motif. Not unexpectedly, other documented dramatic disorder-order transitions which accompany specific DNA binding, like those seen with the N-terminal arms of λ Repressor and the homeodomains, involve very basic polypeptides which protrude from the globular cores.

My CD spectroscopic results show that this coil-to-helix folding transition is required for specific binding of b/HLH/Z proteins to their targets. However, they also show that formation of the basic region helix alone does not result in high-affinity binding. Most reported CD spectroscopic experiments on these proteins, as well as recent proteolytic protection footprinting experiments (S. Cohen, A.R.F., S.K. Burley, and B. Chait, in preparation) confirm that, in general, the basic regions of b/HLH and b/HLH/Z proteins are more disordered when the protein is associated with non-specific DNA than when it is bound to specific DNA. However, results such as those of Figs. 6B and 9C show that non-specific association can result in CD spectroscopically measured helical contents as high or even higher than that resulting from specific association. It has to be borne in mind that what is being measured in these single-concentration spectra is not an affinity, but just the average degree of alignment of the carbonyl dipoles of the polypeptide backbone. Although CD is a useful measurement of gross helical content or conformational transitions, because of its sensitivity to slight environmental perturbations (such as the apparent increase in helical content of b/HLH proteins when measured in buffers containing glycerol; Section 3.1) the actual helical content values should be regarded as approximations. Von Hippel and colleagues (Johnson *et al.*, 1994) have found that in very low salt (10 mM sodium phosphate) even short (twenty residues) alanine-lysine peptides of a sequence conducive to formation of amphipathic helices would mimic the basic region

folding transition when challenged with double stranded DNA (non-specific DNA, by definition). (An interesting experiment in this system was suggested by P.B. Sigler of Yale University: what happens if the same peptides are challenged with double stranded RNA?) In summary, depending on its exact context, the basic regions of b/HLH and b/HLH/Z proteins might adopt various degrees of helicity, independent of the strength of its interaction with the local DNA sequence it is being presented with. The correlation between high helical content and high affinity is strictly true only for some sets of specific and non-specific sequences. It remains to be seen whether atypical basic regions recognizing non-E-box sequences also adopt the fully α-helical conformation seen in the available structures. My CD spectroscopic results would suggest that to be the case (Figs. 23A - 23C).

Because of the entropic penalty, coupling a folding transition to binding, specially if the folding transition occurs fully only upon specific binding, perforce reduces the maximum affinity achievable. Furthermore, if the folding transition does not accompany non-specific binding, this also results in lowered specificity. In the case of the N-terminal arm of λ Repressor, the entropy loss resulting from ordering upon DNA-binding is clearly more than offset by the formation of enthalpically favorable contacts and/or release of solvent or counterions, since presence of the arm increases affinity by 8000-fold (Clarke *et al.*, 1991). Since the folding transition creates most of the DNA-protein interface of b/HLH and b/HLH/Z proteins, there is no reason to believe that it increases affinity or specificity. At a minimum, the folding transition should result in lowered affinity. Yet the b/HLH, b/HLH/Z and b/Z proteins have not been selected against in the evolution of eukaryotes. Why would lower affinity be of biological value? It is possible that lower affinity might be advantageous biologically because it will result in a greater fraction of these proteins being dissociated from the DNA, ready to homo- or hetero-dimerize as the populations of potential dimerization partners change in the nucleus concomitant with homeostasis or

development. Johnson *et al.* (1994) point out that all three families of proteins have as a key element of their biological function the ability to dimerize selectively, and that this ability is central to their biological function. In this light, the perplexing conservation of a solvent-exposed hydrophobic residue in the basic regions of b/HLH/Z proteins noted in the pervious section could be thought to be a means of tuning the affinity of these proteins for DNA to a lower value than the maximum attainable, allowing them to come off the DNA and recombine with other helix-loop-helix proteins.

Structural Basis of Dimerization Specificity The four-helix bundle provides b/HLH and b/HLH/Z proteins with a common three-dimensional scaffold for homo- and heterodimerization. This conserved dimer interface could be made to associate selectively either by allowing certain homo- and heterodimeric combinations more efficient packing of the hydrophobic core, or by decorating the solvent-exposed surfaces of the protomers with side-chains poised to make favorable or unfavorable electrostatic interactions. The results of site-directed mutagenesis studies of the MyoD/E12 heterodimer by Shirakata *et al.* (1993) demonstrate that the higher stability of the heterodimer over the MyoD homodimer is determined by a few salt-bridges and hydrogen bonds that are present in the surface of the four-helix bundle of the high-stability heterodimer. Similar results have been obtained with the Myc/Max heterodimer by Davis and Halazonetis (1993). Not surprising, the leucine zipper coiled-coil stabilizes the dimer interaction only if it is in the proper helical registration relative to H2 (Beckmann and Kadesch, 1991). It is possible to drive selective dimerization of b/HLH/Z *in vitro* and *in vivo* by introducing a very strong leucine zipper dimerization interface (Amati *et al.*, 1993a). This was exploited by these authors to show the requirement of heterodimerization with Max for the oncogenic activity of cMyc to become apparent.

The result of a naïve model building exercise where the Max leucine zipper was used to generate a model of the cMyc/Max heterodimeric zipper is shown on Figs. 25C and 25D. In these figures, the hydrophobic residues at the coiled coil interface were omitted. Figure 25C shows that there are no inter-helical ion pairs present in the Max structure. All charged side-chains are either paired with other charged residues on the same helix or contacting symmetry-related DNA molecules in the crystal. Substitution of the cMyc amino acid sequence into the Max backbone atomic coordinates shows that two favorable interhelical ion pairs might form in the heterodimer. cMyc has a glutamate at position 78 (in the Max numbering scheme) which might ion-pair with Lysine 78 of Max; the cMyc glutamate at position 85 might pair with Max Histidine 81 (incidentally, burial of this histidine also might destabilize the Max homodimer). Two arginines at positions 92 and 102 of cMyc probably destabilize the cMyc homodimer, driving it to heterodimerize. The model shown in Fig. 25D assumes no rearrangement of the packing angles or the hydrophobic interface of the coiled coil. Structure determination will be required to assess its validity, but the figure can constitute a starting point for biochemical experiments.

The Apparent Redundancy of Dimerization Interfaces Human USF binds DNA with nanomolar affinity, either in its full-length form (Pognonec and Roeder, 1991), as the intact b/HLH/Z segment, or as a truncated DNA binding domain that contains the b/HLH and the four most N-terminal helical turns of the coiled-coil (Gregor *et al.*, 1990). Removal of these four helical turns giving the minimal b/HLH consensus results in a protein that retains wild-type DNA specificity, but has an equilibrium dissociation constant as much as 1000-fold higher than the intact b/HLH/Z segment (Ferré-D'Amaré *et al.*, 1994). In collaboration with D. Goss (Hunter College, City University of New York) we have recently determined the dissociation constants of the USF b/HLH/Z and b/HLH constructs for sDNA and nsDNA by employing the intrinsic fluorescence of Tryptophan 218. The dissociation constants were 4.5×10^{-10}, 9.0×10^{-6}, 2.1×10^{-6}, and 3.1×10^{-6} M

for b/HLH/Z and b/HLH binding to sDNA and nsDNA, respectively (D. Goss, A.R.F., S.K. Burley, unpublished). Interestingly b/HLH/Z showed two binding modes; the first one (at 4:1 protein-DNA stoichiometry) had a dissociation constant 4.5×10^{-10}; the second one (at 4:2 stoichiometry) was 8.3×10^{-9} M. Clearly, deletion of the leucine zipper greatly weakens the dimerization of USF. NMR spectroscopy on the USF b/HLH construct in the absence of DNA shows that even at 0.5 mM concentration, the core of the protein is not ordered: deletion of the leucine zipper has resulted in the creation of a "molten globule" (D. Cowburn, A.R.F., and S.K. Burley, unpublished). This disorder seen in the apo-protein might carry over to some extent to the protein-DNA complex, resulting in the plasticity exhibited by this protein in the crystalline state (Section 3.4).

In the USF b/HLH cocrystal structure (Section 3.4), the basic regions make canonical Class-B contacts with the DNA, but the packing of the four-helix bundle is altered because the axis of H2 is 20° closer to the helical axis of the DNA than in the other three crystal structures (Fig. 22D). The high degree of conservation of the residues that make up the hydrophobic core of the HLH and the observation that the naturally zipperless proteins (MyoD and E47) have the same "high" packing angle seen in Max (Section 4.1) imply that the intact b/HLH/Z domain of USF has a Max-like packing. Deletion of the USF leucine zipper region resulted in abnormal helix packing and concurrent weak dimerization. Thus it appears that this b/HLH/Z protein has evolved to require the coiled-coil to brace the HLH in the orientation for optimal packing of the hydrophobic core and for achieving high-affinity dimerization. The existence of zipperless but otherwise highly homologous b/HLH proteins that dimerize efficiently implies that very limited changes in the composition of the hydrophobic core of the four-helix bundle should allow USF to dispense with the coiled-coil. The discovery of a well-dimerizing, zipperless USF in a sea-urchin (Kozlowski *et al.*, 1991) showed that this is indeed the case, and a few amino acid substitutions in H2 are sufficient to achieve stable dimerization in the absence of a C-terminal coiled-coil (Fig. 2).

The fact that the USF truncation mentioned above having the b/HLH consensus and two leucine repeats (four helical turns) of the zipper binds DNA tightly (Gregor *et al.*, 1990) is consistent with this idea.

The comparison of human and echinoderm USF, my results with USF, those of Davis and Halazonetis (1993) in the Myc/Max system, and those of Beckmann and Kadesch (1991) with TFE3/USF chimeras imply that b/HLH/Z proteins have evolved to require the leucine zipper to achieve proper dimerization affinity and specificity. The structural requirement for the coiled-coil also prevents heterodimerization with b/HLH proteins, providing a first-order selectivity "filter". Given the large number of transcription factors with b/HLH and b/HLH/Z motifs potentially present at the same time in the same cell, and the biological importance of selective dimerization, it is not surprising that both proper packing of the hydrophobic core and superficial polar interactions dictate physiologically appropriate homo- and heterodimerization.

Tetramerization and Biological Function b/HLH/Z proteins also form higher order oligomers. At physiologic intranuclear concentration, the USF b/HLH/Z domain exists as a bivalent homotetramer (Ferré-D'Amaré *et al.*, 1994; this work). Independent simultaneous binding to two separate segments of DNA containing the cognate sequence was demonstrated by a bimolecular ligation experiment (Section 3.1). This property is abolished by deletion of the leucine zipper, probably because of alteration in the packing of the HLH. Tetramerization has also been demonstrated for Myc b/HLH/Z by chemical cross-linking (Dang *et al.*, 1989), and for TFEB by glycerol gradient sedimentation and gel-filtration chromatography (Fisher *et al.*, 1991). The biological significance of these higher-order oligomers is not yet understood. It is certainly possible that some proteins, such as the yeast b/HLH/Z centromere binding protein CBF1 (Cai and Davis, 1990) which is a structural component of the kinetochore and is required for chromosome stability, and

the human b/HLH protein CNEP-B which associates with the centromere and forms an oligomeric complex with two physically separate DNA molecules (Muro *et al.*, 1992) play structural roles in the cell nucleus, by virtue of their ability to assemble into bivalent tetramers.

The discovery that human USF can bind to and activate transcription from the adenovirus major late promoter not only through the upstream E-box but also through the initiator element (Du *et al.*, 1993), led to the suggestion (Ferré-D'Amaré *et al.*, 1994) that the bivalent protein tetramer might be mediating DNA looping, helping bring other upstream activators to the vicinity of the general transcription initiation machinery bound at the transcription start site. I mentioned a few paragraphs before that USF b/HLH/Z exhibits two modes of binding, with the first dimer of a presumably tetrameric unit having *higher* affinity for DNA than the second; that is, the tetramer exhibits anticooperativity. This could have the interesting biological consequence of favoring looping with E-boxes near the first one over distant E-boxes, because the proximity translates into an effectively higher local concentration.

In an intriguing recent publication, Ziff and colleagues (Li *et al.*, 1994) demonstrated that cMyc, in addition to activating transcription (as a heterodimer with Max) through canonical Class-B E-boxes, can repress transcription of susceptible genes through the initiator element. This repression required the E-box, an intact b/HLH/Z domain in cMyc, and a stretch of amino acids shared by the Myc oncoproteins outside the DNA-binding and dimerization domain these authors call the Myc box II (MBII). Deletion of MBII had no effect on transcriptional activation by cMyc, but abolished its repressive function. These authors suggest that while proteins like USF may activate genes through both the initiator and upstream E-boxes and possibly induce cell differentiation, when induced, cMyc could compete for binding to the initiator, and by repressing transcription of genes whose

products are important for cell differentiation, bring about blockage of cell differentiation, a step in the path to cellular transformation. It is clear that cMyc/Max heterotetramers could simultaneously bind to both elements of susceptible genes, and that this simultaneous binding might stabilize what is possibly a weak binding (by analogy to the weak binding of USF to the initiator observed by Roy *et al.*, 1991) to isolated initiator elements by this heteromeric b/HLH/Z complex.

Recently, Artandi *et al.* (1994) have documented another instance of b/HLH/Z protein function that can be rationalized in terms of tetramer formation. In their investigation of the transcriptional activation of V_H promoters by the IgH enhancer, these investigators found that the presence of functional E-boxes in both promoter and enhancer elements, and TFE3 protein with an intact b/HLH/Z domain, were required for the synergistic effect to be observed. Furthermore, these authors demonstrated tetramerization *in vitro* of TFE3, and most interestingly, that binding of TFE3 to the promoter E-box enhanced recruitment of TFE3 to the enhancer E-box. It is tempting to interpret these results as supportive of DNA looping across ~ 2 kbp by a b/HLH/Z domain-mediated tetramer.

Association with Atypical DNA Sequences It was pointed out in the Introduction (Section 1.3) that there are a number of b/HLH and b/HLH/Z proteins with highly conserved HLH and Z regions and somewhat divergent basic regions, which recognize DNA sequences unrelated to the E-box such as SREBP-1, which has a b/HLH/Z domain closely related to that of Max, yet has a target DNA does not resemble an E-box: ATCAC(C/G)CCA(C/T). The only significant difference between the SREBP-1 basic region and the consensus (Fig. 2) is replacement of the all-important Arginine 35 with tyrosine. This substitution may permit the conserved Glutamate 32 to participate in completely different protein-DNA interactions. Other divergent b/HLH proteins shown in Fig. 2 are the Dioxin Receptor and its DNA-binding partner, Arnt. This heterodimer

recognizes the sequence GCGTGA. I have already mentioned the b/HLH protein Hairy, which has a proline in its, probably kinked, basic region. Structural investigation of these atypical helix-loop-helix proteins should provide insights into the complicated problem of DNA recognition by members of this family of transcription factors. An interesting possibility is that in the context of other transcription factors bound to nearby sites in the same promoter, b/HLH and b/HLH/Z proteins might bind to, and exert their biological functions, through sequences to which they bind with less than maximum affinity. The biologically productive interaction of USF with the LCR and the initiator elements, and of cMyc/Max with the initiator could be illustrations of such a phenomenon.

4.4 Future Directions

We originally set out to determine the three-dimensional structure of the b/HLH/Z domain, and thereby provide an intellectual scaffold for understanding how this conserved domain is necessary and sufficient for both specific dimerization and DNA-binding. The combination of my structure determinations and the large accumulated body of results from biochemical and genetic investigation by many researchers, provide us today, at the end of 1994, with a comprehensive, if mostly qualitative, understanding of the mode of action of these proteins. Future biochemical and structural work on this family of eukaryotic transcription factors must address the quantitative aspects of DNA binding and dimerization energetics and kinetics, and the structural and energetic bases for specific heterodimerization. I discovered that structurally, the helix-loop-helix domain is comparatively simple. The observation of higher order oligomerization and DNA-induced protein folding suggests that functionally, these proteins are quite complex, even *in vitro*. An important undertaking will be the exploration of the function of helix-loop-helix proteins inside cells, of how dimerization, tetramerization, and DNA-binding specificities

of myriad b/HLH, b/HLH/Z, and HLH proteins, and their interactions with components of the transcriptional apparatus, result in homeostasis and development.

Appendix

Use of Dynamic Light Scattering to Assess Crystallizability of Macromolecules and Macromolecular Assemblies

Introduction Today, crystallization is the rate limiting step in macromolecular structure determination. Successful crystallization is predicated on finding supersaturation conditions where pure preparations will nucleate and crystals grow. The introduction of sparse matrix factorial strategies (Carter and Carter, 1979; Jancarik and Kim, 1991; Doudna *et al.*, 1993) has permitted a systematic search to be made of the enormous parameter space that is relevant for macromolecular crystallization, and advances in recombinant DNA technology have made large amounts of highly purified starting material available. However, covalent purity (assayed commonly by electrophoresis under denaturing conditions, or by mass spectrometry) does not imply monodispersity. It has been found empirically that macromolecules that are monodisperse under "normal" solvent conditions crystallize readily, while those that aggregate randomly and exist as polydisperse

mixtures rarely, if ever, crystallize. Dynamic light scattering can be employed quickly to screen candidate macromolecules or macromolecular assemblages for monodispersity. Then, crystallization trials can be performed only with the molecules that are monodisperse, greatly reducing the effort devoted to projects that are doomed to fail due to the polydisperse nature of the starting materials.

Dynamic Light Scattering Dynamic light scattering (DLS) , also known as photon correlation spectroscopy (PCS) or quasi-elastic light scattering spectroscopy (QELSS), is a technique which allows the translational diffusion coefficient (D_T) of macromolecules to be determined. This is achieved by measuring the time-dependent fluctuations in the intensity of laser light scattered by a solution of the macromolecule. These fluctuations result from the thermal or Brownian motions that molecules undergo in solution. The fluctuations are then analyzed (by constructing an autocorrelation function, the decay of which is related to D_T) and the resulting value of diffusion coefficient can be used to estimate the equivalent radius of gyration and molecular weight of the solute (reviewed in Schmitz, 1990). The technique is exquisitely sensitive to higher order aggregation because the intensity of scattered light is proportional to the square of the mass of the scattering particles. The recent availability of compact commercial instrumentation based on solid-state diode lasers has made routine DLS measurement on small (150 microliters of a 1 mg/ml solution, commonly) samples of macromolecules possible. The experimental set-up is briefly described in Section 2.5.

Monodispersity and Crystallizability DLS has been employed to monitor the formation of protein aggregates as a function of solvent conditions and concentration and to develop models of crystal nucleation and growth. The light scattering studies of Kam *et al.* (1978) on lysozyme provided experimental support for a model in which crystallization is a cooperative process involving step-wise addition of molecules to a growing ordered

assembly, whereas precipitation is the result of uncooperative nonspecific aggregation of molecules into an amorphous polymer. Wilson (1990) employed DLS to distinguish solution conditions yielding pre-crystalline or pre-precipitate aggregates and to study the thermodynamics of the process. Employing DLS to monitor solutions of model proteins approaching supersaturation, Mikol *et al.* (1990) found that precipitants which cause aggregation of the protein in undersaturated conditions failed to produce crystals once supersaturation was attained, while precipitants with which the protein remained monodisperse up to the point of nucleation yielded crystals.

A corollary of this last observation is that macromolecules or macromolecular assemblies which exist as monodisperse solutions in a single aggregation state, under solution conditions far from supersaturation, are likely to crystallize, while macromolecules which aggregate randomly are very unlikely to do so. Zulauf and D'Arcy (1992) employed DLS to analyze the aggregation state of fifteen proteins which had been subjected to extensive crystallization trials, and found that in each case that the protein existed as a monodisperse solution, conditions could be found for it to crystallize, while all proteins which existed as mixtures of oligomerization states failed to crystallize.

Additional instances of the correlation between monodispersity and crystallizability are afforded by my studies of helix-loop-helix transcription factors. Upstream Stimulatory Factor (USF) was overexpressed and purified to apparent homogeneity, and shown to be fully active in DNA binding (Section 3.1). Despite considerable effort, no crystals of either USF protein or USF-DNA complexes could be obtained. Inspection of the aggregation state of the 34 kDa protein by DLS immediately revealed a severe aggregation problem (Fig. 7A). The distribution of apparent molecular weights is very broad, irregular, and extends to well beyond two million daltons. A C-terminal truncation of the protein (A/b/HLH) did not improve the biochemical behavior of the protein. However, elimination

of the N-terminal activation domain did result in a dramatic reduction in random aggregation (Fig. 7B). b/HLH/Z contains an intact, fully active, DNA-binding domain and some additional C-terminal residues. When complexed with DNA, and inspected by DLS, it shows a broad unimodal distribution of molecular weights, with no evidence of random aggregation. A further C-terminal truncation to yield the minimal DNA-binding unit, b/HLH, resulted in a construct which exhibited a narrow unimodal distribution of molecular weights, and which readily crystallized upon addition of precipitants. Optimization of the crystallization conditions by conventional sparse matrix approaches eventually yielded diffraction quality specimens, and the structure of this construct complexed with DNA was solved in due course (Section 3.4).

The same methodology of construct screening was employed for Max (Fig. 10A). Max3-113 H6 complexed with DNA exhibited a broad unimodal distribution of molecular weights, when inspected by DLS. Although some possibly crystalline precipitates could be obtained after setting up crystallization trials in a number of conditions, single crystals were not forthcoming. Elimination of ten hexahistidine tag residues to yield Max1-113 aggravated the problem, by producing a substantially aggregating protein. Deletion of eleven more residues, however, produced a protein, Max22-113 which when complexed with DNA exhibited a narrow, unimodal distribution of molecular weights. When the material used for DLS was recovered from the sample cell, concentrated, and precipitants were added, single crystals were obtained. The structure of this complex was discussed in Section 3.3.

Further examples of the correlation between monodispersity and crystallizability have been obtained in a number of different macromolecular systems, in our laboratory and elsewhere. D'Arcy (1994) reported on an extensive DLS/crystallization survey performed at Hoffman-La Roche. Of 66 proteins examined by DLS and subjected to crystallization

trials, 41 crystallized. 44 had narrow unimodal molecular weight distributions when analyzed by DLS at moderate concentrations in "normal" buffers. Of these, 34 gave crystals. 10 proteins had broad unimodal distributions, and only six of these crystallized. Finally, twelve had multimodal distributions, and only one of these gave crystals.

The DLS experiment is quick (a few minutes) and non-destructive; one macromolecular sample can be assayed for monodispersity under a variety of solvent conditions, in the presence of ligands, inhibitors, cofactors, or post-translational modifications, as a function of the redox potential, of partial proteolysis, etc., greatly enhancing, I believe, the likelihood of successful crystallization. Monodispersity of the preparation is also critical for a number of biophysical methods, such as NMR, SAXS, solution neutron scattering, etc. DLS can be very profitably employed as a fast preliminary screening step for samples destined for those techniques as well.

References

Adam, G. and Delbrück, M. (1968). In *Structural chemistry and molecular biology*, eds. A. Rich and N. Davidson. 198-215. San Francisco: Freeman & Co.

Aggarwal, A.K., Rodgers, D.W., Drottar, M., Ptashne, M., and Harrison, S.C. (1988). Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. Science *242* , 899-907.

Amati, B., Brooks, M.W., Levy, N., Littlewood, T.D., Evan, G.I., and Land, H. (1993a). Oncogenic activity of the c-Myc protein requires dimerization with Max. Cell *72* , 233-245.

Amati, B., Littlewood, T.D., Evan, G.I., and H., L. (1993b). The c-Myc protein induces cell cycle progression and apoptosis through dimerization with Max. EMBO J. *12* , 5083-5087.

Anderson, J., Ptashne, M., and Harrison, S.C. (1984). Cocrystals of the DNA-binding domain of phage 434 repressor and a synthetic phage 434 operator. Proc. Natl. Acad. Sci. USA *81* , 1307-1311.

Anderson, J.E., Ptashne, M., and Harrison, S.C. (1987). Structure of the repressor-operator complex of bacteriophage 434. Nature *326* , 846-852.

Anthony-Cahill, S.J., Benfield, P.A., Robert, F., Wasserman, Z.R., Brenner, S.L., Stafford, W.F.I., Altenbach, C., Hubbel, W.L., and DeGrado, W.F. (1992). Molecular characterization of helix-loop-helix peptides. Science *255* , 979-983.

Åqvist, J., Luecke, H., Quiocho, F.A., and Warshel, A. (1991). Dipoles localized at helix termini stabilize charges. Proc. Natl. Acad. Sci. USA *88* , 2026-2030.

Artandi, S.E., Cooper, C., Shrivastava, A., and Calame, K. (1994). The basic helix-loop-helix-zipper domain of TFE3 mediates enhancer-promoter interaction. Mol. Cell. Biol. *14* , 7704-7716.

Aune, K.C. (1978). Molecular weight measurements by sedimentation equilibrium: some common pitfalls and how to avoid them. Meth. Enzymol. *48* , 163-185.

Austin, D.J., Crabtree, G.R., and Schreiber, S.L. (1994). Proximity versus allostery: the role of regulated protein dimerization in biology. Chem. Biol. *1* , 131-136.

Ayer, D.E. and Eisenman, R.N. (1993). A switch from Myc:Max to Mad:Max heterocomplexes accompanies monocyte/macrophage differentiation. Genes Dev. *7* , 2110-2119.

Ayer, D.E., Kretzner, L., and Eisenman, R.N. (1993). Mad: a heterodimeric partner for Max that antagonizes Myc transcriptional activity. Cell *72* , 211-222.

Baldwin, R.L. (1986). Temperature dependence of the hydrophobic interaction in protein folding. Proc. Natl. Acad. Sci. USA *83* , 8069-8072.

Barkley, M.D. and Bourgeois, S. (1978). Repressor recognition of operator and effectors. In *The Operon*, eds. Jeffrey H. Miller and William S. Reznikoff. 177-220. Cold Spring Harbor: Cold Spring Harbor Laboratory.

Beamer, L.J. and Pabo, C.O. (1992). Refined 1.8 Å crystal structure of the λ repressor-operator complex. J. Mol. Biol. *227* , 177-196.

Beckmann, H. and Kadesch, T. (1991). The leucine zipper of TFE3 dictates helix-loop-helix dimerization specificity. Genes Dev. *5* , 1057-1066.

Bendall, A. and Molloy, P.L. (1994). Base preferences for DNA binding by the bHLH-Zip protein USF: effects of $MgCl_2$ on specificity and comparison with binding of Myc family members. Nucl. Ac. Res. *22* , 2801-2810.

Benezra, R., Davis, R.L., Lockshon, D., Turner, D.L., and Wintraub, H. (1990). The protein Id: a negative regulator of helix-loop-helix DNA binding proteins. Cell *61* , 49-59.

Bengal, E., Ransone, L., Scharfmann, R., Dwarki, V.J., Tapscott, S.J., Weintraub, H., and Verma, I.M. (1992). Functional antagonism between c-Jun and MyoD proteins: a direct physical association. Cell *68* , 507-519.

Berg, O.G., Winter, R.B., and von Hippel, P.H. (1981). Diffusion-driven mechanism of protein translocation on nucleic acids. 1. Models and Theory. Biochemistry *20* , 6929-6948.

Billeter, M., Qian, Y.Q., Otting, G., Müller, M., Gehring, W., and Wüthrich, K. (1993). Determination of the nuclear magnetic resonance solution structure of an Antennapedia homeodomain-DNA complex. J. Mol. Biol. *234* , 1084-1097.

Blackwell, T.K., Huang, J., Ma, A., Kretzner, L., Alt, F.W., Eisenman, R.N., and Weintraub, H. (1993). Binding of Myc proteins to canonical and noncanonical DNA sequences. Mol. Cell. Biol. *13* , 5216-5224.

Blackwood, E. and Eisenman, R.N. (1991). Max: A helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc. Science *251* , 1211-1217.

Blackwood, E.M., Lüscher, B., and Eisenmann, R.N. (1992). Myc and Max associate in vivo. Genes Dev. *6* , 71-80.

Blanar, M.A. and Rutter, W.J. (1992). Interaction cloning: identification of a helix-loop-helix zipper protein that interacts with c-Fos. Science *256* , 1014-1018.

Brennan, R.G., Roderick, S.L., Takeda, Y., and Matthews, B.W. (1990). Protein-DNA conformational changes in the crystal structure of a λ Cro-operator complex. Proc. Natl. Acad. Sci. USA *87*, 8165-8169.

Brennan, R.G., Takeda, Y., Kim, J., Anderson, W.F., and Matthews, B.W. (1986). Crystallization of a complex of Cro repressor with a 17 base-pair operator. J. Mol. Biol. *188* , 115-118.

Brenner, S.L., Zlotnick, A., and Stafford, W.F.I. (1993). RecA protein self-assembly II. analytical equilibrium ultracentrifugation studies of the entropy-driven self-association of RecA. J. Mol. Biol. *216* , 949-964.

Bresnick, E.H. and Felsenfeld, G. (1993). Evidence that the transcription factor USF is a component of the human β–globin locus control region heteromeric protein complex. J. Biol. Chem. *268* , 18824-18834.

Brünger, A.T. (1992). X-PLOR v. 3.1 manual. (New Haven: Yale University).

Brünger, A.T., Kuriyan, J., and Karplus, M. (1987). Crystallographic R factor refinement by molecular dynamics. Science *235* , 458-460.

Buckin, V.A., Kankiya, B.I., Rentzeperis, D., and Marky, L.A. (1994). Mg^{2+} recognizes the sequence of DNA through its hydration shell. J. Am. Chem. Soc. *116* , 9423-9429.

Bungert, J., Düring, F., and Seifart, K.H. (1992). Transcription factor eUSF is an essential component of isolated transcription complexes on the duck histone H5 gene and it

mediates interaction of TFIID with a TATA-deficient promoter. J. Mol. Biol. *223* , 885-898.

Burbach, K.M., Poland, A., and Brafield, C.A. (1992). Cloning of the Ah-receptor reveals a distinctive ligand-activated transcription factor. Proc. Natl. Acad. Sci. USA *89* , 8185-8189.

Burgess, R.R. (1991). Use of polyethyleneimine in purification of DNA-binding proteins. Meth. Enzymol. *208* , 3-10.

Cai, M., and Davis, R.W. (1990). Yeast centromere binding protein CBF1, of the helix-loop-helix protein family, is required for chromosome stability and methionine prototrophy. Cell *61* , 437-446.

Canne, L.E., Ferré-D'Amaré, A.R., Burley, S.K., and Kent, S.B.H. (1994). Total chemical synthesis of a unique transcription factor protein: cMyc-Max. (submitted).

Carr, C.S. and Sharp, P.A. (1990). A helix-loop-helix protein related to the immunoglobulin E box-binding proteins. Mol. Cell. Biol. *10* , 4384-4388.

Carter, C.W.J. and Carter, C.W. (1979). Protein crystallization using incomplete factorial experiments. J. Biol. Chem. *254* , 12219-12223.

Carthew, R.W., Chodosh, L.A., and Sharp, P.A. (1985). An RNA polymerase II transcription factor binds to an upstream element in the adenovirus major late promoter. Cell *43* , 439-448.

Chait, B.T, and Kent, S.B.H. (1992). Weighing naked proteins: practical, high-accuracy mass measurements of peptides and proteins. Science *257* , 1885-1894.

Cho, Y., Gorina, S., Jeffrey, P.D., and Pavletich, N.P. (1994). Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. Science *265* , 346-355.

Chothia, C., Levitt, M., and Richardson, D. (1977). Structure of proteins: packing of α-helices and pleated sheets. Proc. Natl. Acad. Sci. USA *74* , 4130-4134.

Chothia, C., Levitt, M., and Richardson, D. (1981). Helix to helix packing in proteins. J. Mol. Biol. *145* , 215-250.

Chuprina, V.P., Rullmann, J.A.C., Lamerichs, R.M.J.N., van Boom, J.H., Boelens, R., and Kaptein, R. (1993). Structure of the complex of lac repressor headpiece and an 11 base-pair half-operator determined by nuclear magnetic resonance spectroscopy and restricted molecular dynamics. J. Mol. Biol. *234* , 446-462.

Church, G.M., Sussman, J.L., and Kim, S.-H. (1977). Secondary structural complementarity between DNA and proteins. Proc. Natl. Acad. Sci. USA *74* , 1458-1462.

Clark, K.L., Halay, E.D., Lai, E., and Burley, S.K. (1993). Cocrystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. Nature *364* , 412-420.

Clarke, N.D., Beamer, L.J., Goldberg, H.R., Berkower, C., and Pabo, C.O. (1991). The DNA binding arm of λ repressor: critical contacts from a flexible region. Science *254* , 267-270.

Cohen, C. and Parry, D.A.D. (1990). α-helical coiled-coils and bundles: how to design an α-helical protein. Proteins Struct. Funct. Genet. *7* , 1-15.

Creamer, T.P. and Rose, G.D. (1992). Side-chain entropy opposes α-helix formation but rationalizes experimentally determined helix-forming propensities. Proc. Natl. Acad. Sci. USA *89* , 5937-5941.

Crick, F.H.C. (1953a). The Fourier transform of a coiled-coil. Acta Cryst. *6* , 685-689.

Crick, F.H.C. (1953b). The packing of α-helices: simple coiled-coils. Acta Cryst. *6* , 689-697.

Cubellis, M.V., Marino, G., Mayol, L., Piccialli, G., and Sannia, G. (1985). Use of fast protein liquid chromatography for the purification of synthetic oligonucleotides. J. Chromat. *329* , 406-414.

Dang, C.V., Dolde, C., Gillison, M.L., and Kato, G.J. (1992). Discrimination between related DNA sites by a single amino-acid residue of Myc-related basic-helix-loop-helix proteins. Proc. Nat. Acad. Sci. U.S.A. *89* , 599-602.

Dang, C.V., McGuire, M., Buckmire, M., and Lee, W.M.F. (1989). Involvement of the "leucine zipper" region in oligomerization and transforming activity of human c-Myc protein. Nature *337* , 664-666.

D'Arcy, A. (1994). Crystallizing proteins - a rational approach? Acta Cryst. *D50* , 469-471.

Davis, L.J. and Halazonetis, T.D. (1993). Both the helix-loop-helix and the leucine zipper motifs of c-Myc contribute to its dimerization specificity with Max. Oncogene *8* , 125-132.

Davis, R.L., Cheng, P.-F., Lassar, A.B., and Weintraub, H. (1990). The MyoD DNA binding domain contains a recognition code for muscle-specific gene activation. Cell *60* , 733-746.

DiGabriele, A.D., Sanderson, M.R., and Steitz, T.A. (1989). Crystal lattice packing is important in determining the bend of a DNA dodecamer containing an adenine tract. Proc. Natl. Acad. Sci. USA *86* , 1816-1820.

Dill, K.A. (1990). Dominant forces in protein folding. Biochemistry *29* , 7133-7155.

Dong, Q., Blatter, E.E., Ebright, Y.W., Bister, K., and Ebright, R.H. (1994). Identification of amino acid-base contacts in the Myc-DNA complex by site-specific bromouracil-mediated photocrosslinking. EMBO J. *13* , 200-204.

Doudna, J., Grosshans, C., Gooding, A., and Kundrot, C.E. (1993). Crystallization of ribozymes and small RNA motifs by a sparse matrix approach. Proc. Natl. Acad. Sci. USA *90* , 7829-7833.

Drenth, J. (1994). Principles of protein X-ray crystallography. New York: Springer-Verlag.

Du, H., Roy, A.L., and Roeder, R.G. (1993). Human transcription factor USF stimulates transcription through the initiator elements of the HIV-1 and the Ad-ML promoters. EMBO J. *12* , 501-511.

Durchschlag, H. (1986). Specific volumes of biological macromolecules and some other molecules of biological interest. In *Thermodynamic data for biochemistry and biotechnology*, ed. Hans-Jürgen Hinz. 45-128. Berlin: Springer-Verlag.

Edelhoch, H. (1967). Spectroscopic determination of tryptophan and tyrosine in proteins. Biochemistry *6* , 1948-1954.

Eisenberg, D. and McLachlan, A.D. (1986). Solvation energy in protein folding and binding. Nature *319* , 199-203.

Eliason, J.L., Weiss, M.A., and Ptashne, M. (1985). NH_2-terminal arm of phage λ repressor contributes energy and specificity to repressor binding and determines the effects of operator mutations. Proc. Natl. Acad. Sci. USA *82* , 2339-2343.

Ellenberger, T., Fass, D., Arnaud, M., and Harrison, S.C. (1994). Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. Genes Dev. *8* , 970-980.

Ellenberger, T.E., Brandl, C.J., Struhl, K., and Harrison, S.C. (1992). The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted a helices: crystal structure of the protein-DNA complex. Cell *71* , 1223-1237.

Ellis, H.M., Spann, D.R., and Posakony, J.W. (1990). Extramacrochaetae, a negative regulator of sensory organ development in Drosophila, defines a new class of helix-loop-helix proteins. Cell *61* , 27-38.

Fairall, L., Schwabe, J.W., Chapman, L., Finch, J.T., and Rhodes, D. (1993). The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. Nature *366* , 483-487.

Fairman, R., Beran-Steed, R.K., Anthony-Cahill, S.J., Lear, J.D., Stafford, W.F.I., DeGrado, W.F., Benfield, P.A., and Brenner, S.L. (1993). Multiple oligomeric states regulate the DNA-binding of helix-loop-helix peptides. Proc. Natl. Acad. Sci. USA *90* , 10429-10433.

Feldman, T., Alex, R., Suckow, J., Dildrop, R., Kisters-Woike, G., and Müller-Hill, B. (1993). Single exchanges of amino acids in the basic region change the specificity of N-Myc. Nucl. Acids Res. *21* , 5050-5058.

Felsenfeld, G. (1992). Chromatin as an essential part of the transcriptional mechanism. Nature *355* , 219-224.

Ferré-D'Amaré, A.R., Pognonec, P., Roeder, R.G., and Burley, S.K. (1994). Structure and function of the b/HLH/Z domain of USF. EMBO J. *13* , 180-189.

Ferré-D'Amaré, A.R., Prendergast, G.C., Ziff, E.B., and Burley, S.K. (1993). Recognition by Max of its cognate DNA through a dimeric b/HLH/Z domain. Nature *363* , 38-45.

Fickert, R. and Müller-Hill, B. (1992). How lac repressor finds lac operator in vitro. J. Mol. Biol. *226* , 59-68.

Finkelstein, A.V. and Janin, J. (1989). The price of lost freedom: entropy of bimolecular complex formation. Prot. Eng. *3* , 1-3.

Fisher, D.E., Parent, L.A., and Sharp, P.A. (1992). Myc/Max and other helix-loop-helix/leucine zipper proteins bend DNA toward the minor groove. Proc. Natl. Acad. Sci. USA *89* , 11779-11783.

Fisher, F. and Goding, C.R. (1992). Single amino acid substitutions alter helix-loop-helix protein specificity for bases flanking the core CANNTG motif. EMBO J. *11* , 4103-4109.

Fujimoto, B.S. and Schurr, M. (1990). Dependence of the torsional rigidity of DNA on base composition. Nature *344* , 175-178.

Gannon, F., O'Hare, K., Perrin, F., LePennec, J.P., Benoist, C., Cochet, M., Brethnach, R., Royal, A., Garapin, A., Cami, B., and Chambon, P. (1979). Organisation and sequences at the 5' end of a cloned complete ovalbumin gene. Nature *278* , 428-434.

Gibson, T.J., Thompson, J.D., and Abagyan, R.A. (1993). Proposed structure for the DNA-binding domain of the helix-loop-helix family of eukaryotic gene regulatory proteins. Protein Eng. *6* , 41-50.

Gilbert, W. and Maxam, A. (1973). The nucleotide sequence of the lac operator. Proc. Natl. Acad. Sci. USA *70* , 3581-3584.

Gilbert, W. and Müller-Hill, B. (1966). Isolation of the lac repressor. Proc. Natl. Acad. Sci. USA *56* , 1891-1898.

Gilbert, W. and Müller-Hill, B. (1967). The lac operator is DNA. Proc. Natl. Acad. Sci. USA *58* , 2415-2421.

Gilson, M.K., Sharp, K.A., and Honig, B.H. (1988). Calculating the electrostatic potential of molecules in solution: method and error assessment. J. Comp. Chem. *9* , 327-335.

Gregor, P.D., Sawadogo, M., and Roeder, R.G. (1990). The adenovirus major late transcription factor USF is a memeber of the helix-loop-helix group of regulatory proteins and binds to DNA as a dimer. Genes Dev. *4* , 1730-1740.

Gruner, S.L. (1994). X-ray detectors for macromolecular crystallography. Curr. Op. Struct. Biol. *4* , 765-769.

Ha, J.-H., Spolar, R.S., and Record, T.M.J. (1989). Role of the hydrophobic effect in stability of site-specific protein-DNA complexes. J. Mol. Biol. *209* , 801-816.

Halazonetis, T.D. and Kandil, A.N. (1992). Predicted structural similarities of the DNA bionding domains of c-Myc and endonuclease Eco RI. Science *255* , 464-466.

Harrington, R.E. (1978). Opticohydrodynamic propterties of high-molecular weight DNA. III. The effects of NaCl concentration. Biopolymers *17* , 919-936.

Harris, N.L., Presnell, S.R., and Cohen, F.E. (1994). Four helix bundle diversity in globular proteins. J. Mol. Biol. *236* , 1356-1368.

Harrison, S. (1991). A structural taxonomy of DNA-binding domains. Nature *353* , 715-719.

He, J.J. and Quiocho, F.A. (1993). Dominant role of local dipoles in stabilizing uncompensated charges on a sulfate sequestered in a periplasmic transport protein. Protein Sci. *2* , 1643-1647.

Hegde, R.S., Grossman, S.R., Laimins, L.A., and Sigler, P.B. (1992). Crystal structure at 1.7Å of the bovine papillomavirus-1 E2 DNA-binding domain bound to its DNA target. Nature *359* , 505-512.

Hodgkinson, C.A., Moore, K.J., Nagayama, A., Steingrímsson, E., Copeland, N.G., Jenkins, N.A., and Arnheiter, H. (1993). Mutations at the mouse microphthalmia locus are associated with defects in a gene encoding a novel basic-helix-loop-helix-zipper protein. Cell *74* , 395-404.

Hoffman, E.C., Reyes, H., Chu, F.-F., Sander, F., Conley, L.H., Brooks, B.A., and Hankinson, O. (1991). Cloning of a factor required for activity of the Ah (Dioxin) receptor. Science *252* , 954-958.

Hogan, M.E. and Austin, R.H. (1987). Importance of DNA stiffness in protein-DNA binding specificity. Nature *329* , 263-266.

Hol, W.G.J. (1985). The role of the α-helix dipole in protein function and structure. Prog. Biophys. Molec. Biol. *45* , 149-195.

Hsu, H.-L., Huang, L., Tsan, J.T., Funk, W., Wright, W.E., Hu, J.-S., Kingston, R.E., and Baer, R. (1994). Preferred sequences for DNA recognition by the TAL1 helix-loop-helix proteins. Mol. Cell. Biol. *14* , 1256-1265.

Hua, X., Yokoyama, C., Wu, J., Briggs, M.R., Brown, M.S., Goldstein, J.L., and Wang, X. (1993). SREBP-2, a second basic-helix-loop-helix-leucine zipper protein that stimulates transcription by binding to a sterol regulatory element. Proc. Natl. Acad. Sci. USA *90* , 11603-11607.

Hurlburt, B.K. and Yanofsky, C. (1992). The NH_2-terminal arms of trp repressor participate in repressor-operator association. Nucl.Acids Res. *20* , 337-341.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. *3* , 318-356.

Jancarik, J. and Kim, S.-H. (1991). Sparse matrix sampling: a screening method for crystallization of proteins. J. Appl. Cryst. *24* , 409-411.

Janin, J. and Chothia, C. (1990). The structure of protein-protein recognition sites. J. Biol. Chem. *265* , 16027-16030.

Jeltsch, A., Alves, J., Wolves, H., Maass, G., and Pingoud, A. (1994). Pausing of the restriction endonuclease EcoRI during linear diffusion on DNA. Biochemistry *33* , 10215-10219.

Jencks, W.P. (1981). On the attribution and additivity of binding energies. Proc. Natl. Acad. Sci. USA *78* , 4046-4050.

Joachimiak, A. and Sigler, P.B. (1991). Crystallization of protein-DNA complexes. Meth. Enzymol. *208* , 82-99.

Johnson, N.P., Lindstrom, J., Baase, W.A., and von Hippel, P.H. (1994). Double-standed DNA templates can induce alpha-helical conformation in peptides containing lysine and alanine: functional implications for leucine zipper and helix-loop-helix transcription factors. Proc. Natl. Acad. Sci. USA *91* , 4840-4844.

Jones, T.A., Zou, J.Y., Cowan, S.W., and Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Cryst. *A47* , 110-119.

Jordan, S.R. and Pabo, C.O. (1988). Structure of the lambda complex at 2.5 Å resolution: details of the repressor-operator interactions. Science *242* , 893.

Kabsch, W., Mannherz, H.G., Suck, D., Pai, E.F., and Holmes, K.C. (1990). Atomic structure of the actin:DNase I complex. Nature *347* , 37-44.

Kam, Z., Shore, H.B., and Feher, G. (1978). On the crystallization of proteins. J. Mol. Biol. *123* , 539-555.

Kao-Huang, Y., Revzin, A., Butler, A.P., O'Conner, P., Noble, D.W., and von Hippel, P.H. (1977). Nonspecific DNA bindig of genome-regulating proteins as a biological

control mechanism: measurement of DNA-bound Escherichia coli lac repressor in vivo. Proc. Natl. Acad. Sci. USA *74* , 4228-4232.

Karplus, P.A. and Faerman, C. (1994). Ordered water in macromolecular structure. Curr. Op. Struct. Biol. *4* , 770-776.

Kaulen, H., Pognonec, P., Gregor, P.D., and Roeder, R.G. (1991). The Xenopus B1 factor is closely related to the mammalian activator USF and is implicated in the developmental regulation of TFIIIA gene expression. Mol. Cell. Biol. *11* , 412-424.

Kim, J.L. and Burley, S.K. (1994). 1.9 Å resolution refined structure of TBP recognizing the minor groove of TATAAAAG. Nature Struct. Biol. *1* , 638-653.

Kim, J.L., Nikolov, D.B., and Burley, S.K. (1993a). Cocrystal structure of TBP recognizing the minor groove of a TATA element. Nature *365* , 520-527.

Kim, Y., Geiger, J.H., Hahn, S., and Sigler, P.B. (1993b). Crystal Structure of a yeast TBP/TATA-box complex. Nature *356* , 512-520.

Kirschbaum, B.J., Pognonec, P., and Roeder, R.G. (1992). Definition of the transcriptional activation domain of recombinant 43-kilodalton USF. Mol. Cell. Biol. *12* , 5094-5101.

Kissinger, C.R., Liu, B., Martin-Blanco, E., Kornberg, T.B., and Pabo, C.O. (1990). Crystal structure of an Engrailed homeodomain-DNA complex at 2.8Å resolution: a framework for understanding homeodomain-DNA interactions. Cell *63* , 579-590.

Klein, C. and Struhl, K. (1994). Increased recruitment of TATA-Binding Protein to the promoter by transcriptional activation domains in vivo. Science *266* , 280-282.

Klemm, J.D., Rould, M.A., Aurora, R., Herr, W., and Pabo, C.O. (1994). Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding domains. Cell *77* , 21-32.

Koblan, K.S. and Ackers, G.K. (1991a). Cooperative protein-DNA interactions: effects of KCl on λ cI binding to O_r. Biochemistry *30* , 7822-7827.

Koblan, K.S. and Ackers, G.K. (1991b). Energetics of subunit dimerization in bacteiphage λ cI repressor: linkage to protons temperature and KCl. Biochemistry *30* , 7817-7821.

Koshland Jr., D.E. (1958). Application of a theory of enzyme specificity to protein synthesis. Proc. Natl. Acad. Sci. USA *44* , 98-114.

Kouzarides, T. and Ziff, E. (1988). The role of the leucine zipper in the fos-jun interaction. Nature *336* , 646-651.

Kozlowski, M.T., Gan, L., Venuti, J.M., Sawadogo, M., and Klein, W.H. (1991). Sea urchin USF: a helix-loop-helix protein active in embryonic ectoderm cells. Developmental Biol. *148* , 625-630.

Kraulis, P.J. (1991). Molscript: a program to produce both detailed and schematic plots of protein structures. J. Appl. Cryst. *24* , 946-950.

Kretzner, L., Blackwood, E.M., and Eisenman, R.N. (1992). Myc and Max proteins possess distinct transcriptional activities. Nature *359* , 426-429.

Kuntz, I.D. and Kauzmann, W. (1974). Hydration of proteins and polypeptides. Adv. Prot. Chem. *28* , 239-345.

Ladbury, J.E., Wright, J.G., Sturtevant, J.M., and Sigler, P.B. (1994). A thermodynamic study of the trp repressor-operator interaction. J. Mol. Biol. *238* , 669-681.

Laemmli, U.K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature *227* , 680-685.

Landschultz, W.H., Johnson, P.F., and McKnight, S.L. (1988). The leucine zipper: a hypothetical structure common to a new class of DNA-binding proteins. Science *240* , 1759-1564.

Laskowski, R.J., Macarthur, M.W., Moss, D.S., and Thornton, J.M. (1993). PROCHECK: a program to check stereochemical quality of protein structures. J. Appl. Cryst. *26* , 283-290.

Lassar, A.B., Buskin, J.N., Lockson, D., Davis, R.L., Apone, S., Hauschka, S.D., and Weintraub, H. (1989). MyoD is a sequence-specific DNA binding protein requiring a region myc homology to bind to the muscle creatine kinase enhancer. Cell *58* , 823-831.

Lavery, R. and Sklenar, H. (1989). Defining the structure of irregular nucleic acids: conventions and principles. J. Biomol. Struct. Dynamics *6* , 655-667.

Lawson, C.L. and Carey, J. (1993). Tandem binding in crystals of a trp repressor/operator half-site complex. Nature *366* , 178.

Lee, B. and Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. *55* , 379-400.

Lehming, N., Sartorius, J., Kisters-Woike, B., von Wilcken-Bergmann, B., and Müller-Hill, B. (1990). Mutant lac repressors with new specificities hint at rules for protein-DNA recognition. EMBO J. *9* , 615-621.

Leirmo, S., Harrison, C., Cayley, S.D., Burgess, R.R., and Record Jr., T.M. (1987). Replacement of potassium chloride by potassium glutamate dramtically enhances protein-DNA interactions in vitro. Biochemistry *26* , 2095-2101.

Li, L.-H., Nerlov, C., Prendergast, G., MacGregor, D., and Ziff, E.B. (1994). c-Myc represses transcription in vivo by a novel mechanism dependent on the initiator element and Myc box II. EMBO J. *13* , 4070-4079.

Liao, X. and Butow, R.A. (1993). RTG1 and RTG2: two yeast genes required for a novel path of communication from mitochondria to nucleus. Cell *72* , 61-72.

Luisi, B.F., Xu, W.X., Otwinowski, Z., Feedman, L.P., Yamamoto, K.R., and Sigler, P.B. (1991). Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. Nature *352* , 497-505.

Luzzati, P.V. (1952). Traitement statistique des erreurs dans la determination des structures cristallines. Acta Cryst. *5* , 802-810.

Ma, P.C.M., Rould, M.A., Weintraub, H., and Pabo, C.O. (1994). Crystal structure of MyoD bHLH domain-DNA complex: perspectives on DNA recognition and implications for transcriptional activation. Cell *77* , 451-459.

Manning, G.S. (1978). The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. Quart. Rev. Biophys. *11* , 179-246.

Mark, A.E. and van Gunsteren, W.F. (1994). Decomposition of the free energy of a system in terms of specific interactions implications for theoretical and experimental studies. J. Mol. Biol. *240* , 167-176.

Marmorstein, R., Carey, M., Ptashne, M., and Harrison, S.C. (1992). DNA recognition by Gal4: structure of a protein-DNA complex. Nature *356* , 408-414.

Marmorstein, R. and Harrison, S.C. (1994). Crystal structure of a PPR1-DNA complex: DNA recognition by proteins containing a Zn2Cys6 binuclear cluster. Genes Dev. *8* , 2504-2512.

Martin, J.L., Waksman, G., Bardwell, J.C.A., Beckwith, J., and Kuriyan, J. (1993). Crystallization of DsbA, an Escherichia coli protein required for disulphide bond formation in vivo. J. Mol. Biol. *230* , 1097-1100.

Matthews, B.W. (1968). Solvent content of protein crystals. J. Mol. Biol. *33* , 491-497.

McPherson, A. (1990). Current approaches to macromolecular crystallization. European J. Biochem. *189* , 1-23.

Meisterernst, M., Horikoshi, M., and Roeder, R.G. (1990). Recombinant yeast TFIID, a general transcription factor, mediates activation by the gene-specific factor USF in a chromatin assembly assay. Proc. Natl. Acad. Sci. USA *87* , 9153-9157.

Mikol, V., Hirsch, E., and Giegé, R. (1990). Diagnostic of precipitant for biomacromolecule crystallization by quasi-elastic light-scattering. J. Mol. Biol. *213* , 187-195.

Miller, S., Lesk, A.M., Janin, J., and Chothia, C. (1987). The accessible surface area and stability of oligomeric proteins. Nature *328* , 834-836.

Misra, V.K., Hecht, J.L., Sharp, K.A., Friedman, R.A., Honig, B. (1994). Salt effects on protein-DNA interactions. J. Mol. Biol. *238*, 264-280.

Mitchell, P.J. and Tjian, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science *245* , 371-378.

Miyamoto, N.G., Moncollin, V., Egly, J.M., and Chambon, P. (1985). Specific interaction between a transcription factor and the upstream element of the adenovirus-2 major late promoter. EMBO J. *4* , 3563-3570.

Mondragón, A. and Harrison, S.C. (1991). The phage 434 Cro/OR1 complex at 2.5Å resolution. J. Mol. Biol. *219* , 321-334.

Muir, T.W. and Kent, S.B.H. (1993). The chemical synthesis of proteins. Curr. Op. Biotech. *4* , 420-427.

Mukherjee, B., Morgenbesser, S.D., and DePinho, R.A. (1992). Myc family oncoproteins function through a common pathway to transform normal cells in culture: cross-interference by Max and trans-acting dominant mutants. Genes Dev. *6* , 1480-1492.

Muro, Y., Masumoto, H., Yoda, K., Nozaki, N., Ohashi, M., and Okazaki, T. (1992). Centromere protein B assembles human centromeric α-satellite DNA at the 17-bp sequence, CENP-B Box. J. Cell. Biol. *116* , 585-596.

Murphy, K.P., Xie, D., Thompson, K.S., Amzel, M., and Freire, E. (1994). Entropy in biological binding processes: estimation of translational entropy loss. Proteins Struct. Funct. Genet. *18* , 63-67.

Murre, C., McCaw, P.S., and Baltimore, D. (1989a). A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. Cell *56* , 777-783.

Murre, C., McCaw, P.S., Vaessin, H., Caudy, M., Jan, L.Y., Jan, Y.N., Cabrera, C.V., Buskin, J.N., Hauschka, S.D., Lassar, A.B., Weintraub, H., and Baltimore, D. (1989b). Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. Cell *58* , 537-544.

Murzin, A.G. and Finkelstein, A.V. (1988). General architecture of the α-helical globule. J. Mol. Biol. *204* , 749-769.

Neuhold, L.A. and Wold, B. (1993). HLH forced dimers: tethering MyoD to E47 generates a dominant positive myogenic factor insulated from negative regulation by Id. Cell *74* , 1033-1042.

Ogawa, N. and Oshima, Y. (1990). Functional domains of a positive regulatory protein, Pho4, for transcriptional control of the phosphatase regulon in Saccharomyces cerevisiae. Mol. Cell. Biol. *10* , 2224-2236.

Ohsako, S., Hyer, J., Panganiban, G., Oliver, I., and Caudy, M. (1994). Hairy function as a DNA binding HLH repressor of Drosophila sensory organ development. Genes Dev. (in the press).

Omichinski, J.G., Clore, G.M., Schaad, O., Felsenfeld, G., Trainor, C., Appella, E., Stahl, S., and Gronenborn, A.M. (1993). NMR structure of a specific DNA complex of Zn-containing DNA binding domain of GATA-1. Science *261* , 438-446.

Otwinowski, Z., Schevitz, R.W., Zhang, R.-G., Lawson, C.L., Joachimiak, A., Marmorstein, R.Q., Luisi, B.F., and Sigler, P.B. (1988). Crystal structure of trp repressor/operator complex at atomic resolution. Nature *335* , 321-329.

Pabo, C.O. and Sauer, R.T. (1992). Transcription factors: structural families and principles of DNA recognition. Annu. Rev. Biochem. *61* , 1053-1095.

Parker, C.S. and Roeder, R.G. (1977). Selective and accurate transcription of the Xenopus laevis 5S RNA genes in isolated chromatin by purified RNA polymerase III. Proc. Natl. Acad. Sci. USA *74* , 44-48.

Pauling, L., Corey, R.B., and Branson, H.R. (1951). The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. USA *37* , 205-211.

Pavletich, N.P. and Pabo, C.O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. Science *252* , 809-817.

Pavletich, N.P. and Pabo, C.O. (1993). Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. Science *261* , 1701-1707.

Pickett, S.D. and Sternberg, M.J.E. (1993). Empirical scale of side-chain conformational entropy in protein folding. J. Mol. Biol. *231* , 825-839.

Poellinger, L., Göttlicher, M., and Gustafsson, J.-Å. (1992). The dioxin and peroxisome proliferator-activated receptors: nuclear receptors in search of endogenous ligands. Trends Pharm. Sci. *13* , 241-245.

Pognonec, P., Kato, H., and Roeder, R.G. (1992). The helix-loop-helix/leucine repeat transcription factor USF can be functinally regulated in a redox-dependent manner. J. Biol. Chem. *267* , 24563-24567.

Pognonec, P., Kato, H., Sumimoto, H., Kretzschmar, M., and Roeder, R.G. (1991). A quick procedure for purification of functional recombinant proteins over-expressed in E. coli. Nucl. Acids Res. *19* , 6650-6650.

Pognonec, P. and Roeder, R.G. (1991). Recombinant 43-kDa USF binds to DNA and activates transcription in a manner indistinguishable from that of natural 43/44-kDa USF. Mol. Cell. Biol. *11* , 5125-5136.

Prendergast, G.C., Hopewell, R., Gorham, B., and Ziff, E.B. (1992). Biphasic effect of Max on Myc transformation activity and dependence on N- and C-terminal Max functions. Genes Dev. *6* , 2429-2439.

Prendergast, G.C., Lawe, D., and Ziff, E.B. (1991). Association of Myn, the murine homolog of Max with c-Myc stiumulates methylation-sensitive DNA binding and Ras cotransformation. Cell *65* , 395-407.

Prendergast, G.C. and Ziff, E.B. (1989). DNA-binding motif. Nature *341* , 392.

Presnell, S.R. and Cohen, F.E. (1989). Topological distribution of four-α-helix bundles. Proc. Natl. Acad. Sci. USA *86* , 6592-6596.

Provencher, S.W. and Glöckner, J. (1981). Estimation of globular protein secondary structure from circular dichroism. Biochemistry *20* , 33-37.

Ptashne, M. (1967a). Isolation of the λ phage repressor. Proc. Natl. Acad. Sci. USA *57* , 306-313.

Ptashne, M. (1967b). Specific binding of the λ phage repressor to DNA. Nature *214* , 232-234.

Ptashne, M. (1988). How eukaryotic transcriptional activators work. Nature *335* , 683-689.

Puglisi, J.D. and Tinoco Jr., I. (1989). Absorbance melting curves of RNA. Meth. Enzymol. *180* , 304-.

Qian, Y.Q., Otting, G., Billeter, M., Müller, M., Gehring, W., and Wüthrich, K. (1993a). Nuclear magnetic resonance spectroscopy of a DNA complex with the uniformly ^{13}C labeled Antennapedia homeodomain and strcuture determination of the DNA-bound homeodomain. J. Mol. Biol. *234* , 1070-1083.

Qian, Y.Q., Otting, G., and Wüthrich, K. (1993b). NMR detection of hydration water in the intermolecular interface of a protein-DNA complex. J. Am. Chem. Soc. *115* , 1189-1190.

Ramachandran, G.N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of Polypeptide Chain Configuration. J. Mol. Biol. *7* , 95-99.

Raumann, B.E., Rould, M.A., Pabo, C.O., and Sauer, R.T. (1994). DNA recognition by β–sheets in the Arc repressor-operator crystal structure. Nature *367* , 754-757.

Richmond, T., and Richards, F.M. (1978). Packing of α-helices geometrical constraints and contact areas. J. Mol. Biol. *119*, 537-555.

Richter, P.H. and Eigen, M. (1974). Diffusion controlled reaction rates in spheroidal geometry. Application to repressor-operator association and membrane bound enzymes. Biophys. Chem. *2* , 255-263.

Riggs, A.D. and Bourgeois, S. (1968). On the assay, isolation and characterization of the lac repressor. J. Mol. Biol. *34* , 361-364.

Riggs, A.D., Bourgeois, S., and Cohn, M. (1970a). The lac repressor-operator interaction III. Kinetic studies. J. Mol. Biol. *53* , 401-417.

Riggs, A.D., Bourgeois, S., Newby, R.F., and Cohn, M. (1968). DNA binding of the lac repressor. J. Mol. Biol. *34* , 365-368.

Riggs, A.D., Newby, R.F., and Bourgeois, S. (1970b). lac repressor-operator interaction II. Effect of galactosides and other ligands. J. Mol. Biol. *51* , 303-314.

Riggs, A.D., Suzuki, H., and Bourgeois, S. (1970c). lac repressor-operator interaction I. Equilibrium studies. J. Mol. Biol *48* , 67-83.

Roeder, R.G. (1991). The complexities of eukaryotic transcription initiation: regulation of preinitiation complex assembly. Trends Biochem. Sci. *16* , 402-408.

Roeder, R.G. and Rutter, W.J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. Nature *224* , 234-237.

Roy, A.L., Malik, S., Meisterernst, M., and Roeder, R.G. (1993). An alternative pathway for transcription initiation involving TFII-I. Nature *365* , 355-359.

Roy, A.L., Meisterernst, M., Pognonec, P., and Roeder, R.G. (1991). Cooperative interaction of an initiator-binding transcription initiation factor and the helix-loop-helix activator USF. Nature *354* , 245-248.

Rushlow, C.A., Hogan, A., Pinchin, S.M., Howe, K.M., Lardelli, M., and Ish-Horowicz, D. (1989). The Drosophila hairy protein acts in both segmentation and bristle patterining and shows homology to N-myc. EMBO J. *8* , 3095-3103.

Saenger, W. (1984). Principles of Nucleic Acid Structure. New York: Springer-Verlag.

Sakabe, N. (1991) X-ray diffraction data collection system for modern protein crystallography with a Weissenberg camera and an imaging plate using synchrotron radiation. Nucl. Instr. Meth. Phys. Res. *A303* , 448-463.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular cloning a laboratory manual. Second Ed. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.

Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and structural meaning of sequence alignment. Proteins Struc. Funct. Genet. *9* , 56-68.

Sato, M., Yamamoto, M., Imada, K., Katsube, Y., Tanaka, N., and Higashi, T. (1992). A high-speed data-collection system for large-unit-cell crystals using an imaging plate as a detector. J. Appl. Cryst. *25* , 348-357.

Sawadogo, M. (1988). Multiple forms of the human gene-specific transcription factor USF. II. DNA binding properties and transcriptional activity of the purified HeLa USF. J. Biol. Chem. *263* , 11994-12001.

Sawadogo, M. and Roeder, R.G. (1985). Interaction of a gene-specific transcription factor with the adenovirus major late promoter upstream of the TATA box region. Cell *43* , 165-175.

Sawadogo, M., Van Dyke, M.W., Gregor, P.D., and Roeder, R.G. (1988). Multiple forms of the human gene-specific transcription factor USF. I. Complete purification and identification of USF from HeLa cell nuclei. J. Biol. Chem. *263* , 11985-11993.

Schägger, H. and von Jagow, G. (1987). Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa. Anal. Biochem. *166* , 368-379.

Scheraga, H.A. and Mandelkern, L. (1953). Consideration of the hydrodynamic properties of proteins. J. Am. Chem. Soc. *75* , 179-184.

Schleif, R. (1992). DNA looping. Annu. Rev. Biochem. *61* , 199-223.

Schneider, T.D., Stromo, G.D., Gold, L., and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. J. Mol. Biol. *188* , 415-431.

Schmitz, K.H. (1990). An introduction to dynamic light scattering by macromolecules. New York: Academic Press.

Schultz, S.C., Shields, G.C., and Steitz, T.A. (1991). Crystal structure of a CAP-DNA complex: the DNA is bent by 90°. Science *253* , 1001-1007.

Schwabe, J.W.R., Chapman, L., Finch, J.T., and Rhodes, D. (1993). The crystal structure of the estrogen receptor DNA-binding doamin bound to DNA: how receptrors discriminate between their response elements. Cell *75* , 567-579.

Seeman, N.C., Rosenberg, J.M., and Rich, A. (1976). Sequence-specific recognition of doulbe helical nucleic acids by proteins. Proc. Natl. Acad. Sci. USA *73* , 804-808.

Senear, D.F. and Batey, R. (1991). Comparison of operator-specific and nonspecific DNA binding of the λ cI repressor: [KCl] and pH effects. Biochemistry *30* , 6677-6688.

Seo, J. and Cohen, C. (1993). Pitch diversity in α-helical coiled coils. Proteins Struct. Funct. Genet. *15*, 223-234.

Seto, E., Shi, Y., and Shenk, T. (1991). YY1 is an initiator sequence-binding protein that directs and activates transcription in vitro. Nature *354* , 241-245.

Shirakata, M., Friedman, F.K., Wei, Q., and Paterson, B.M. (1993). Dimerization specificity of myogenic helix-loop-helix DNA-binding factors directed by nonconserved hydrophilic residues. Genes Devel. *7* , 2456-2470.

Sigler, P.B. (1993). The surprising chemistry of specific protein nuclei acid interfaces. In *Proceedings of the Robert A. Welch Foundation 37th conference on chemical research 40 years of the DNA double helix.* 63-76. Houston: The Robert W. Welch Foundation.

Sirito, M., Lin, Q., Maity, T., and Sawadogo, M. (1994). Ubiquitous expression of the 43- and 44-kDa forms of transcription factor USF in mammalian cells. Nucl. Acids Res. *22* , 427-433.

Sklar, V.E.F. and Roeder, R.G. (1977). Transcription of specific genes in isolated nuclei by exogenous RNA polymerases. Cell *10* , 405-411.

Smale, S.T. and Baltimore, D. (1989). The "initiator" as a transcription element. Cell *57* , 103-113.

Solomon, D.L.C., Amati, B., and Land, H. (1993). Distinct DNA binding preferences for the c-Myc/Max and Max/Max dimers. Nucl. Acids Res. *21* , 5372-5376.

Somers, W.S. and Phillips, S.E. (1992). Crystal structure of the Met repressor-operator complex at 2.8 Å resolution reveals DNA recognition by β-strands. Nature *359* , 387-393.

Spolar, R.S., Ha, J.-H., and Record, T.M.J. (1989). Hydrophobic effect in protein folding and other noncovalent processes involving proteins. Proc. Natl. Acad. Sci. USA *86* , 8382-8385.

Spolar, R.S. and Record, T.M. (1994). Coupling of local folding to site-specific binding of proteins to DNA. Science *263* , 777-784.

Squires, C.L., Lee, F.D., and Yanofsky, C. (1975). Interaction of the trp repressor and RNA polymerase with the trp operon. J. Mol. Biol. *92* , 93-111.

Starovasnik, M.A., Blackwell, T.K., Laue, T.M., Weintraub, H., and Klevit, R.E. (1992). Folding topology of the disulfide-bonded dimeric DNA-binding domain of the myogenic determination factor MyoD. Biochemistry *31* , 9891-9903.

Steingrímsson, E., Moore, K.J., Lamoreux, M.L., Ferré-D'Amaré, A.R., Burley, S.K., Sanders-Zimring, D.C., Skow, L.C., Hodgkinson, C.A., Arnheiter, H., Copeland, N.G., and Jenkins, N.A. (1994). Molecular basis of mouse microphthalmia (mi) mutations helps explain their developmental and phenotypic consequences. Nature Genetics *8* , 256-263.

Steitz, T.A. (1990). Structural studies of protein-nucleic acid interaction: the sources of sequence-specific binding. Quart. Rev. Biophys. *23* , 105-180.

Stuart, D.I. and Jones, E.Y. (1993). Weissenberg data collection for macromolecular crystallography. Curr. Op. Struct. Biol. *3* , 737-740.

Studier, F.W., Rosenberg, A.H., Dunn, J.J., Dubendroff, J.W. 1990. Use of T7 RNA polymerase to direct expression of cloned genes. Meth. Enzymol. *185*, 60-89.

Stura, E.A. and Wilson, I.A. (1990). Analytical and production seeding techniques. Methods *1* , 38-49.

Sturtevant, J.M. (1977). Heat capacity and entropy changes in processes involving proteins. Proc. Natl. Acad. Sci. USA*74* , 2236-2240.

Sun, X.-H. and Baltimore, D. (1991). An inhibitory domain of E12 transcription factor prevents DNA binding in E12 homodimers but not in E12 heterodimers. Cell *64* , 459-470.

Sundralingam, M. and Rao, S.T., eds. (1975). *Structure and conformation of nucleic acids and protein-nucleic acid interactions*. Baltimore: University Park Press.

Swanson, S.M. (1988). Effective resolution of macromolecular X-ray diffraction data. Acta Cryst. *A44*, 437-442.

Takeda, Y., Ross, P.D., and Mudd, C.P. (1992). Thermodynamics of Cro protein-DNA interactions. Proc. Natl. Acad. Sci. USA *89* , 8180-8184.

Terwilliger, T.C. and Eisenberg, D. (1983). Unbiased three-dimensional refinement of heavy-atom parameters by correlation of origin-removed patterson functions. Acta Cryst. *A39* , 813-817.

Vinson, C.R. and Garcia, K.C. (1992). Molecular model for DNA recognition by the family of basic-helix-loop-helix-zipper proteins. New Biol. *4* , 396-403.

Vinson, C.R., Sigler, P.B., and Mcknight, S.L. (1989). Scissors-grip model for DNA recognition by a family of leucine zipper proteins. Science *246* , 911-916.

von Hippel, P.H. and Berg, O.G. (1986). On the specificity of DNA-protein interactions. Proc. Natl. Acad. Sci. USA *83* , 1608-1612.

Voronova, A. and Baltimore, D. (1990). Mutations that disrupt DNA binding and dimer formation in the E47 helix-loop-helix protein map to distinct domains. Proc. Natl. Acad. Sci. USA *87* , 4722-4726.

Wadman, I., Li, J., Bash, R.O., Forster, A., Osada, H., Rabbitts, T.H., and Baer, R. (1994). Specific in vivo association between the bHLH and LIM proteins implicated in human T cell leukemia. EMBO J. *13* , 4831-4839.

Weber, P.C. and Salemme, F.R. (1980). Structural and functional diversity in 4-α-helical proteins. Nature *287* , 82-84.

Wechsler, D.S. and Dang, C.V. (1992). Opposite orientations of DNA bending by c-Myc and Max. Proc. Natl. Acad. Sci. USA *89* , 7635-7639.

Wechsler, D.S., Papoulas, O., Dang, C.V., and Kingston, R.E. (1994). Differential binding of c-Myc and Max to nucleosomal DNA. Mol. Cell. Biol. *14* , 4097-4107.

Weiss, M.A., Ellenberger, T., Wobbe, R.C., Lee, J.P., Harrison, S.C., and Struhl, K. (1990). Folding transition in the DNA-binding domain of GCN4 on specific binding to DNA. Nature *347* , 575-578.

Wilson, D., Sheng, G., Lecuit, T., Dostatni, N., and Desplan, C. (1993). Cooperative dimerization of Paired class homeo domains on DNA. Genes Dev. *7* , 2120-2134.

Wilson, W.W. (1990). Monitoring crystallization experiments using dynamic light scattering: assaying and monitoring protein crystallization in solution. Methods *1* , 110-117.

Winter, B., Braun, T., and Arnold, H.-H. (1992). Co-operativity of functional domains in the muscle-specific transcription factor Myf-5. EMBO J. *11* , 1843-1855.

Winter, R.B., Berg, O.G., and von Hippel, P.H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor-operator interaction: Kinetic measurements and conclusions. Biochemistry *20* , 6961-6977.

Winter, R.B. and von Hippel, P.H. (1981). Diffusion-driven mechanisms of protein translocation on nucleic acids. 2. The Eschericha coli repressor-operator interaction: equilibrium measurements. Biochemistry *20* , 6948-6960.

Wolberger, C., Dong, Y., Ptashne, M., and Harrison, S.C. (1988). Structure of a phage 434 Cro/DNA complex. Nature *335* , 789-795.

Wolberger, C., Vershon, A.K., Liu, B., Johnson, A.D., and Pabo, C. (1991). Crystal structure of a MAT α2 homeodomain-operator complex suggests a general model for homeodomain-DNA interactions. Cell *67* , 517-528.

Wolffe, A. (1994). The transcription of chromatin templates. Curr. Op. Genetics Dev. *4* , 245-254.

Workman, J.L., Roeder, R.G., and Kingston, R.E. (1990). An upstream transcription factor, USF (MLTF), facilitates the formation of preinitiation complexes during in vitro chromatin assembly. EMBO J. *9* , 1299-1308.

Yokoyama, C., Wang, X., Briggs, M.R., Admon, A., Wu, J., Hua, X., Goldstein, J.L., and Brown, M.S. (1993). SREBP-1, a basic-helix-loop-helix-leucine zipper protein that controls transcription of the low-density lipoprotein receptor gene. Cell *75* , 187-197.

Zawel, L. and Reinberg, D. (1993). Initiation of transcription by RNA polymerase II: a multi-step process. Prog. Nucl. Ac. Res. Molec. Biol. *44* , 67-108.

Zeelen, J.P. and Wierenga, R.K. (1992). The growth of yeast thiolase crystals using a polyacrylamide gel as dialysis membrne. J. Crystal Growth *122* , 194-198.

Zervos, A.S., Gyuris, J., and Brent, R. (1993). Mxi1, a protein that specifically interacts with max to bind myc-max recognition sites. Cell *72* , 223-232.

Zhang, K.Y.J. and Main, P. (1990). The use of Sayre's equation with solvent flatteing and histogram matching for phase extension and refinement of protein structures. Acta Cryst. *A46* , 377-381.

Ziff, E.B. and Evans, R.M. (1978). Coincidence of the promoter and capped 5' terminus of RNA from the adenovirus 2 major late transcription unit. Cell *15* , 1463-1475.

Zimm, B.H. and Bragg, J.K. (1959). Theory of the phase transition between helix and random coil in polypeptide chains. J. Chem. Phys. *31* , 526-535.

Zulauf, M. and D'Arcy, A. (1992). Light scattering of proteins as a criterion for crystallization. J. Cryst. Growth *122* , 102-106.

End