Evidence That the DNA Binding Specificity of Winged Helix Proteins Is Mediated by a Structural Change in the Amino Acid Sequence Adjacent to the Principal DNA Binding Helix[†]

Ian Marsden,[‡] Yuan Chen,[§] Changwen Jin,[‡] and Xiubei Liao*,[‡]

Department of Biochemistry, University of Illinois at Chicago, 1819 W. Polk Street, Chicago, Illinois 60612, and Division of Immunology, Beckman Research Institute of the City of Hope, 1450 E. Duarte Road, Duarte, California 91010

Received June 24, 1997; Revised Manuscript Received August 27, 1997[⊗]

ABSTRACT: We present the first structural evidence supporting the hypothesis that the binding specificity of the winged helix DNA binding motif is mediated by residues adjacent to the α-helix (H3), the moiety which is primarily involved in the interaction with DNA. Using NMR to determine secondary structural elements of a winged helix family member, Genesis (formerly HFH-2), and comparing these with those found in the X-ray crystal structure of the HNF-3γ/DNA complex [Clark, K. L., Halay, E. D., Lai, E., & Burley, S. K. (1993) *Nature 364*, 412−420], we show that the major differences observed occur for H3 and the region immediately prior to this. H3 in Genesis is slightly shorter than in HNF-3γ and, in addition, we observe an extra small helix (H4) in the region between H2 and H3 which is not found in the HNF-3γ/DNA complex. This is significant as it has been shown previously [Overdier, D. G., Porcella, A., & Costa R. H. (1994) *Mol. Cell. Biol. 14*, 2755−2766] that the DNA-binding specificity is influenced by amino acid residues in this region.

One of the most critical steps in transcription regulation is the recognition of a specific DNA sequence by transcription factors leading to enhanced or repressed transcription of target genes (Johnson & McKnight, 1989; Mitchell & Tjian, 1989). This recognition process is mediated by both nonspecific interactions and base specific contacts between amino acid residues of a transcription factor and nucleotide residues of its cognate DNA site. Therefore, an understanding of the sequence specific interactions between transcription factors and their target DNA sites is crucial for the understanding of transcriptional regulation.

It has been shown that transcription factors can be grouped into families based on the conserved DNA recognition motifs that a particular family adopts [reviewed by Pabo and Sauer (1992) and Harrison and Aggarwal (1990)] such as zinc-finger (Miller et al., 1985), homeodomain (Gehring & Hiromi, 1986; Kornberg, 1993), basic-leucine zipper (Landschultz et al., 1988) and the helix-turn-helix motif (Harrison & Aggarwal, 1990; Murre et al., 1989). The DNA binding properties and the structural basis of these motifs for recognizing specific DNA sites are well studied, and many different contact schemes are observed. In many families of DNA binding proteins, the differences in DNA binding specificity are determined by variation of the contact residues on recognition elements; therefore, the substitution of a DNA contact residue within family members will lead to different binding specificity. However, the transcription factor family containing the winged helix binding motif presents an exception.

The winged helix motif has been identified in DNA binding domains of the HNF-3/fkh[1] family and was initially named for the structural features of the first member of this family to be determined, HNF-3γ. It is defined by an approximately 110 amino acid DNA binding domain which is relatively conserved, commonly occurring toward the interior of the full protein, and similar to the motif found in another class of transcription factors, the helix-turn-helix motif. The HNF-3/fkh family was so named for the first proteins identified, i.e., *Drosophila* forkhead homeotic protein (fkh*)* (Weigal & Jäckle, 1990) and rat hepatocyte nuclear factors, HNF-3 (Lai et al., 1990). However, during the last several years, many other hepatocyte nuclear factor/forkhead homologue (HFH) proteins have been identified in various organisms ranging from yeast to human [reviewed by Kaufmann and Knöckel (1996)], and these studies have demonstrated that HFH proteins play important roles in developmental regulation and tissue specific gene regulation (Costa et al., 1989; Costa & Grayson 1991; Liu et al., 1991; Ang & Rossant 1994; Weinstein et al., 1994).

The structure of one family member, HNF-3γ, complexed with double-stranded DNA has been determined by X-ray crystallography (Clark et al., 1993). In this structure, the N-terminal region of HNF-3γ folds into three α-helices resembling the helix-turn-helix motif, where the third helix (H3) is presented into the major groove of the DNA

[†] This work was funded by the National Institutes of Health (Grant GM 52034 to X.L., Grant GM 54190 to Y.C.) and the American Heart Association (Grant 95010100 to X.L.). X.L. and Y.C. are recipients of JFA awards from the American Cancer Society. This study made use of the National Magnetic Resonance Facility at Madison, which is supported by NIH grant RR02301 from the Biomedical Research Technology Program, National Center for Research Resources. Equipment in the facility was purchased with funds from the University of Wisconsin, the NFS Biological Instrumentation Program (DMB-8415048), NFS Academic Research Program (BIR-9214394), NIH Biomedical Research Technology Program (RR02301), NIH Shared Instrumentation Program (RR02781), and the U.S. Department of Agriculture.

* Author to whom correspondence should be addressed.

[‡] University of Illinois at Chicago.

[§] Beckman Research Institute of the City of Hope.

[⊗] Abstract published in *Advance ACS Abstracts*, October 15, 1997.

[1] Abbreviations: HNF, hepatocyte nuclear factor; fkh, forkhead; HFH, hepatocyte nuclear factor/forkhead homologue; NOE, nuclear Overhauser effect; NOESY, nuclear Overhauser enhancement spectroscopy; TOCSY, total correlated spectroscopy; HSQC, heteronuclear single quantum correlation; FREAC, forkhead related activators.

S0006-2960(97)01514-6 CCC: $14.00

providing critical base-specific contacts. A three stranded, twisted, antiparallel β-sheet is also present which incorporates a loop (W1) that interacts with the phosphate backbone. The complex is further stabilized via several nonspecific DNA contacts made by a second wing (W2) at the C-terminus of the domain, hence the term winged helix.

One of the unique properties of HNF-3/fkh homologues is that, although the amino acid sequences in the DNA recognition helix are almost invariable in different family members, HFH proteins show different DNA binding specificity. On the basis of a sequence swap experiment, a hypothesis was proposed to explain how these conserved family members recognize different DNA sequences (Overdier et al., 1994). In this hypothesis, a stretch of 20 amino acid residues immediately adjacent to the DNA recognition helix regulates the relative presentation of this helix in different family members leading to distinct DNA binding specificity for each protein. This 20 amino acid sequence is bounded by the C-terminus of helix 2 and N-terminus of helix 3 and includes the connecting loop between the two. This sequence includes residues which diverge within the HFH family; although, due to conserved residues at distinct positions, the different HFH proteins can be divided into further subgroups. An important question is, therefore, what is the basis for sequence specific binding in this motif given that this amino acid sequence is found to regulate binding specificity. NMR is a powerful method for studying this as localized structural characteristics can be determined accurately on the basis of chemical shifts and NOE patterns.

In this report, we present the first NMR studies on the DNA binding domain of one of these HNF-3/fkh homologues, Genesis (Sutton et al., 1996), and introduce the first structural evidence that the 20 amino acid region adopts alternate structures. Genesis is a 465 residue protein which has been shown to be specifically expressed in embryonic stem cells or their malignant equivalent. Cotransfection experiments revealed also that Genesis is a transcriptional repressor. The DNA binding domain of 100 amino acids between residues 129 and 229 has been shown to be sufficient to retain full binding specificity, and therefore, we chose to study only this section of the protein as the full protein is not amenable to NMR investigation.

MATERIALS AND METHODS

Expression and Purification of the DNA Binding Domain of Genesis. The gene encoding the DNA binding domain of Genesis (Overdier et al., 1994), which in this report will be referred to only as Genesis, was generated by PCR amplification from the rat genomic clone (Clevidence et al., 1993), and the coding sequence used for expression of the protein was confirmed by DNA sequencing (Tabor & Richardson, 1987). The endonuclease recognition sites *Nde*I and *Xho*I were engineered for the cloning of Genesis into pET21b (Novagen) with a 6XHis tag fusion at the C-terminus. Therefore, expressed protein contains the functional DNA binding domain of Genesis with an extra Met at the N-terminus and Leu-Gln-(His)$_6$ at the C-terminus for purification purposes.

The protein was produced in *Escherichia coli* HMS174 by induction of T7 polymerase at midlogarithmic growth through the addition of IPTG to a concentration of 1 mM. Uniformly labeled Genesis was grown in isotopically enriched minimal medium containing 0.6% Na_2HPO_4, 0.3% KH_2PO_4, 0.15% NaCl, 1 mM $MgCl_2$, and 0.1 mM $CaCl_2$, 0.5% Basal Medium Eagle Vitamin Solution (Gibco), and rare element supplement. $(^{15}NH_4)_2SO_4$ (1 g/L) and unlabeled glucose (4 g/L) were used for ^{15}N labeling or 1.5 g/L $(^{15}NH_4)_2SO_4$ and 2 g/L of ^{13}C glucose for double labeling in either H_2O or D_2O. Protein expression was induced at an O.D.$_{600}$ of 0.5 (0.4 in D_2O), and cells were grown for another 12−18 h (24 in D_2O) with the container open to air. Contrary to other reports, we found that the bacteria need not be gradually accustomed to grow in D_2O solution as a small D_2O culture was initially grown to midlog phase and then stored at −70 °C in 15% glycerol solution. This small culture was then used for inoculating larger expression cultures whereupon collected protein was purified using Ni-NTA resin (Qiagen, CA). Since the majority of expressed Genesis was present as insoluble inclusion bodies, a standard procedure using denaturing conditions was performed for extraction [Qiagen Manual (1992), 2nd Ed.). After elution of Genesis in 8 M urea at pH 4.5, the protein was renatured by dialysis against phosphate buffer 100 mM at pH 6.5.

Purified Genesis was exchanged into NMR buffer (50 mM phosphate, pH 6.5, 100 mM NaCl, and 10 mM $Na_2S_2O_4$ in 10% D_2O/90% H_2O) by ultrafiltration. Due to precipitation during dialysis and ultrafiltration, the final concentration for NMR studies was found to be slightly lower than 1 mM. However, this renatured protein appeared to be uniformly folded as judged by a ^{15}N edited HSQC experiment (Figure 1) and also demonstrated fully functional DNA binding toward known DNA binding sites of the protein in gel shift assays.

Genesis−DNA Complex. Formation of the complex was achieved by first exchanging the protein into Tris buffer (10 mM Tris-d^6, 20 mM NaCl, and 10 mM $Na_2S_2O_4$ in 10% D_2O/90% H_2O). Addition of DNA to Genesis initially causes precipitation of the free protein after which complexation then causes this precipitation to redissolve. Therefore, an approximate 1:1 stoichiometry is reached on clarification of the solution.

NMR Spectroscopy. NMR spectra of the free protein were recorded on either a Varian Unity Plus 500 or a Bruker DMX 500 NMR spectrometer, while for the complex, spectra were taken at 600 MHz. Both instruments were equipped with three channels and a pulsed-field-gradient accessory. The NMR data were processed and analyzed using Triad 6.2 software (Tripos, Inc., St. Louis, MO).

The following experiments were recorded on the free protein at 290 K with either ^{15}N, ^{15}N/^{2}H, or ^{13}C/^{15}N labeled samples: 2D [^{15}N]HSQC, 3D ^{1}H-^{15}N-correlated TOCSY and NOESY-HSQC (Fesik & Zuiderweg, 1990), and 3D ^{1}H-^{13}C-correlated NOESY-HSQC. The center frequencies for double resonance experiments were 4.74 ppm (^{1}H) and 118 ppm (^{15}N). Triple resonance experiments HNCA (Kay et al., 1990), HNCACB (Wittekind & Mueller, 1993), and CBCA(CO)NH (Grzesiek & Bax, 1992) for sequential assignment purposes were also recorded. The center frequencies for these triple resonance experiments were 4.74 (^{1}H), 118 (^{15}N), and 44.5 ppm (^{13}C). 3D ^{1}H-^{15}N-correlated NOESY-HSQC spectra were recorded on the ^{15}N/^{2}H-labeled protein (180 ms mixing time) and Genesis−DNA complex (120 ms, 303 K). Sensitivity-enhanced gradient pulse sequences were employed for experiments where magnetization was detected on the amide HN (Muhandiram & Kay, 1994; Palmer et al., 1991).

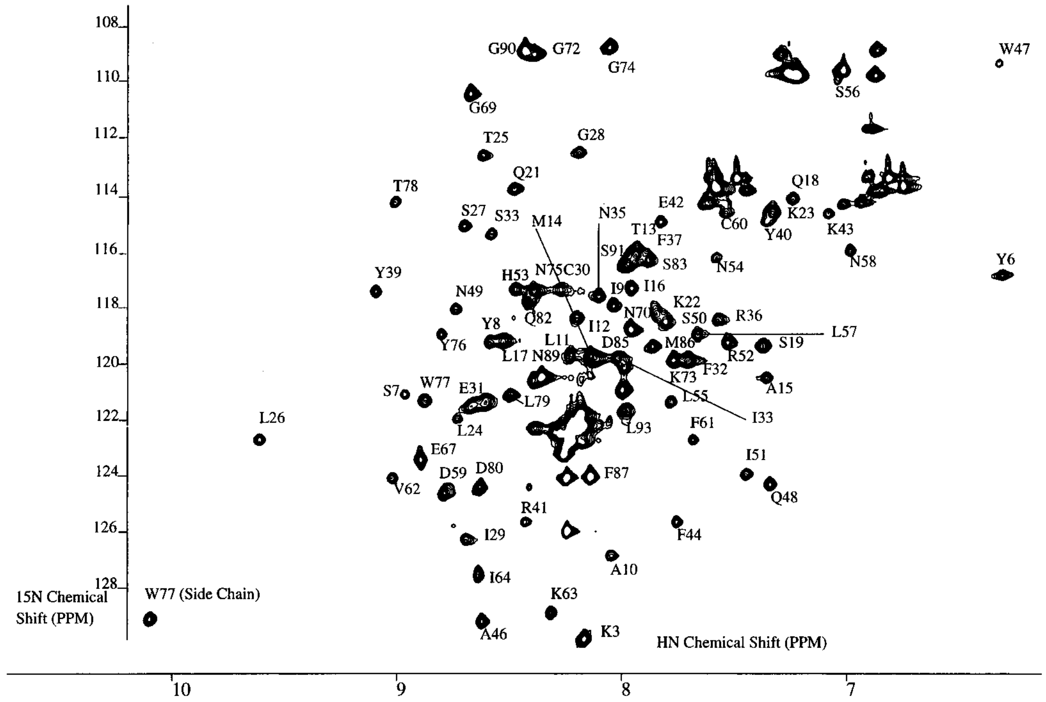

FIGURE 1: ^{1}H-^{15}N HSQC spectrum of Genesis free in solution.

RESULTS

Resonance Assignments. A [^{1}H-^{15}N]HSQC spectrum was initially run to determine that the protein was feasible for further conformational analysis, as were gel shift assays to also determine that the expressed and purified protein showed expected DNA binding activity. The [^{1}H-^{15}N]HSQC spectrum is displayed in Figure 1, which also shows the resonance assignments. These resonance assignments were made with the aid of both triple resonance experiments and 3D ^{15}N resolved NOESY and TOCSY spectra. Starting from distinctive residues such as glycine, small sequence elements were determined through sequential connectivities in triple resonance experiments. Experiments that were utilized in this work include an HNCA, which displayed all assigned C$^\alpha$ resonances, an HNCACB, and a CBCA(CO)NH. From these experiments, the identity of individual amino acids was tentatively determined through characteristic Cα and Cβ chemical shifts, and then these small sequence elements were further assigned to areas in the protein by sequence comparisons. Any ambiguities in sequential connectivities of backbone Cα atoms were first resolved on the basis of Cβ chemical shift comparisons between the CBCA(CO)NH and HNCACB experiment. Then, in the final analysis, sequential assignments were checked against observed NOE cross-peaks, both amide and side chain, in the event that Cβ shifts were also degenerate or missing due to inefficient magnetization transfer. Furthermore, after completing approximately 85% of the backbone assignments, we produced a deuterium-labeled protein which allowed us to complete the assignment procedure and double check existing assignments with a 3D ^{15}N NOESY-HSQC spectrum. This was particularly useful since due to the increased sensitivity on deuteration most $d_{NN}(i, i + 1)$ and many $d_{NN}(i, i + 2)$ correlations were observed. In addition, because of our growth protocol, some amino acids were selectively labeled at known positions thereby leading to a small number of distinctive HN-side chain NOE cross-peaks.

Chemical Shift Index. Through our NMR investigation of Genesis, we have been able to determine the secondary structure through the use of both the chemical shift index and NOE connectivities. These results are summarized in Figures 2, panels a and b, and 3, respectively. The chemical shift index (Wishart et al., 1994a,b) is a quick and simple method to determine regions in the protein which are either α-helix or β-sheet in that Cα and Hα chemical shifts change in relation to a standard value according to the secondary structural elements that an amino acid adopts. The standard values are taken from peptides in a random coil conformation, and it is found empirically that the chemical shifts of Cα and Hα of amino acids involved in α-helices are shifted downfield and upfield, respectively, while those in β-sheets are conversely shifted opposite to this. As can be seen in Figure 2, panels a and b, there are four main regions in the protein which are shifted downfield to a large extent (>1 ppm) and three regions shifted upfield which strongly indicate formation of α-helices and β-sheets, respectively.

Short-Range NOE Connectivities. α-Helices and β-sheet structures typically show very characteristic NOE patterns (Wuthrich, 1986), and therefore, an examination of such patterns will determine those regions in the peptide which assume these secondary elements. For α-helices, sequential correlations between amide protons, $d_{NN}(i, i + 1)$, and in some cases, $d_{NN}(i, i + 2)$ can be observed in addition to correlations from α-protons, $d_{\alpha N}(i, i + 1)$, $d_{\alpha N}(i, i + 2)$, and $d_{\alpha N}(i, i + 3)$. For β-sheet structures, strong cross-strand d_{NN}, $d_{\alpha N}$, and sequential $d_{\alpha N}(i, i + 1)$ NOEs are to be expected with only weak or nonexistent $d_{NN}(i, i + 1)$ correlations. The observed NOEs are summarized in Figure 3 and the cross-strand β-sheet correlations are shown in Figure 4.

On examination of Figure 3, there are four regions in the protein which show characteristic NOEs for α-helices, these mostly occurring in the N-terminal region. In addition, three β-strands forming an antiparallel β-sheet are also observed and this is shown in Figure 4 in more detail. Through the use of a deuterated protein to run a ^{15}N NOESY-HSQC experiment, a significant increase in sensitivity and resolution was achieved for d_{NN} correlations due to a decrease in the number of dipolar relaxation pathways (Venters et al., 1995;

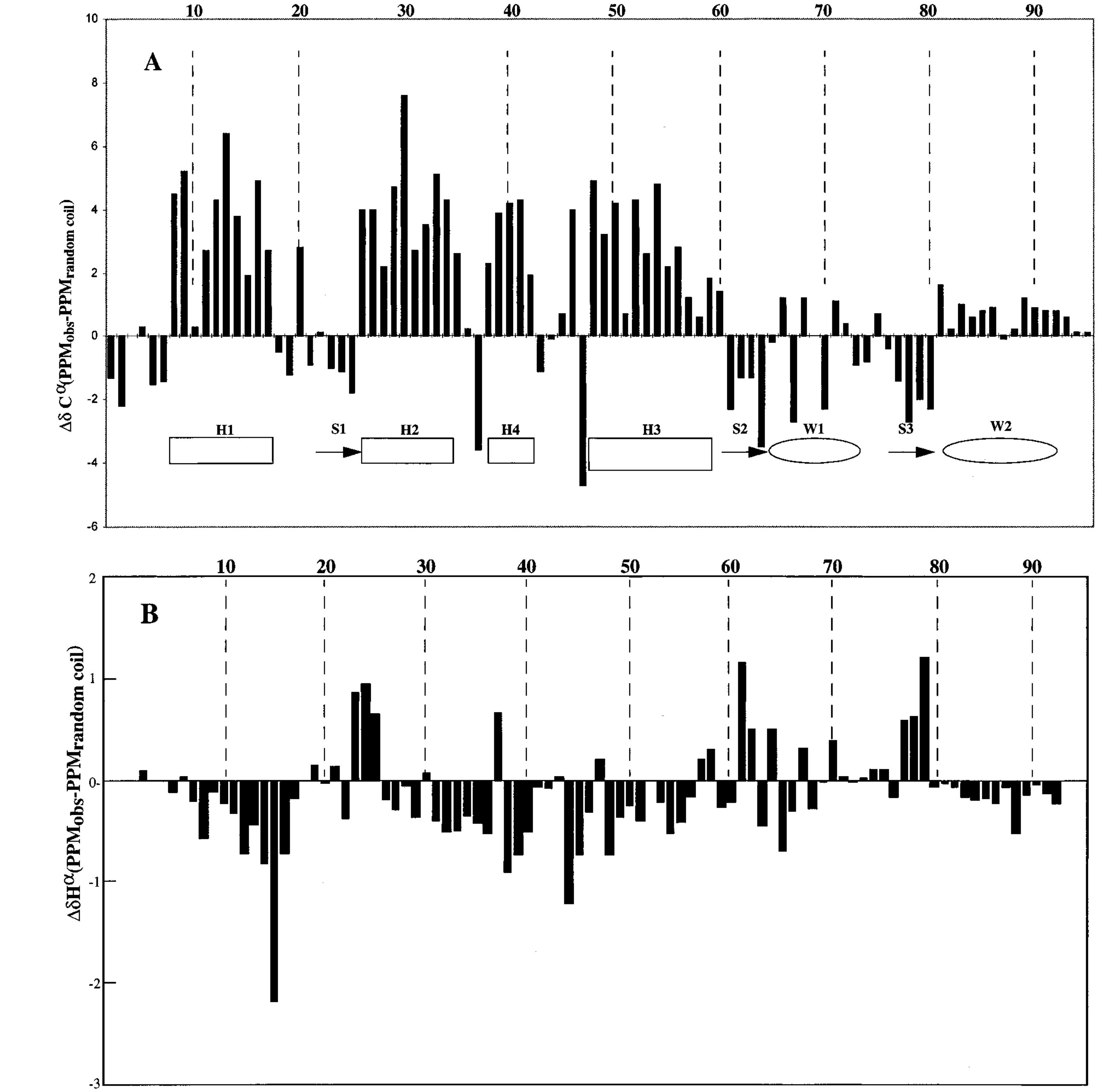

FIGURE 2: (A) Cα secondary shifts [Δδ Cα (PPM$_{obs}$ − PPM$_{random\ coil}$)] plotted as a function of amino acid for Genesis. (B) Hα secondary shifts [Δδ Hα (PPM$_{obs}$ − PPM$_{random\ coil}$)] plotted as a function of amino acid for Genesis. The positions of secondary structural elements predicted by the chemical shift index method (see text for details) are shown underneath.

Farmer & Venters, 1996; Smith et al., 1996; Gardner et al., 1997). Therefore, in Genesis, many $d_{NN}(i, i + 2)$ and even some $d_{NN}(i, i + 3)$ correlations (not shown) were observed also.

Secondary Structure of Genesis. Analysis of the Cα chemical shift differences and NOE pattern demonstrates that the secondary structure is consistent with having one short and three long α-helices, in addition to three relatively short β-strands. The helices are all contained within the N-terminal region between Tyr 8-Gln 18 (H1), Leu 26-Arg 36 (H2), Tyr 39-Lys 43 (H4), and Gln 48-Asn 58 (H3), while the β-sheet strands cover Lys 23-Thr 25 (S1), Phe 61-Ile 64 (S2), and Tyr 76-Asp 80 (S3).

Helix 1 (H1) begins at Tyr 8 as evidenced by the change in the Cα chemical shifts, although the NOE $d_{\alpha N}(i, i + 3)$ data does not correlate with this and suggests that the helix starts at Ala 10 instead. This discrepancy is probably due to the fact that for these residues strong exchange with water is observed in the NOESY spectrum so leading to reduced signal intensities and the loss of small cross-peaks. However, given the large Cα chemical shift value changes for Tyr 8 and Ile 9 (cf. 4.5 and 5.2 ppm, respectively) and strong d_{NN} correlations in this region, Tyr 8 would seem the more likely starting residue. This helix then extends to Gln 18 as shown by strong d_{NN} and $d_{\alpha N}(i, i + 3)$ correlations, before a short random coil is observed leading to the first β-strand (S1). This β-strand is relatively short and only extends three residues between Lys 23 and Thr 25. NOEs for this strand (Figure 4) show that it is in close proximity to the C-terminal residues, Trp 76 and Thr 78. S1 is followed directly by helix 2 (H2) at Leu 26, and this then extends to Arg 36 as suggested by both the Cα and NOE data. H2 is closely followed by helix 4 (H4), which for consistency, we have labeled based on considerations of the HNF-3γ crystal structure which shows only three helices, helix 3 being the principal DNA binding helix. H4 is relatively short in

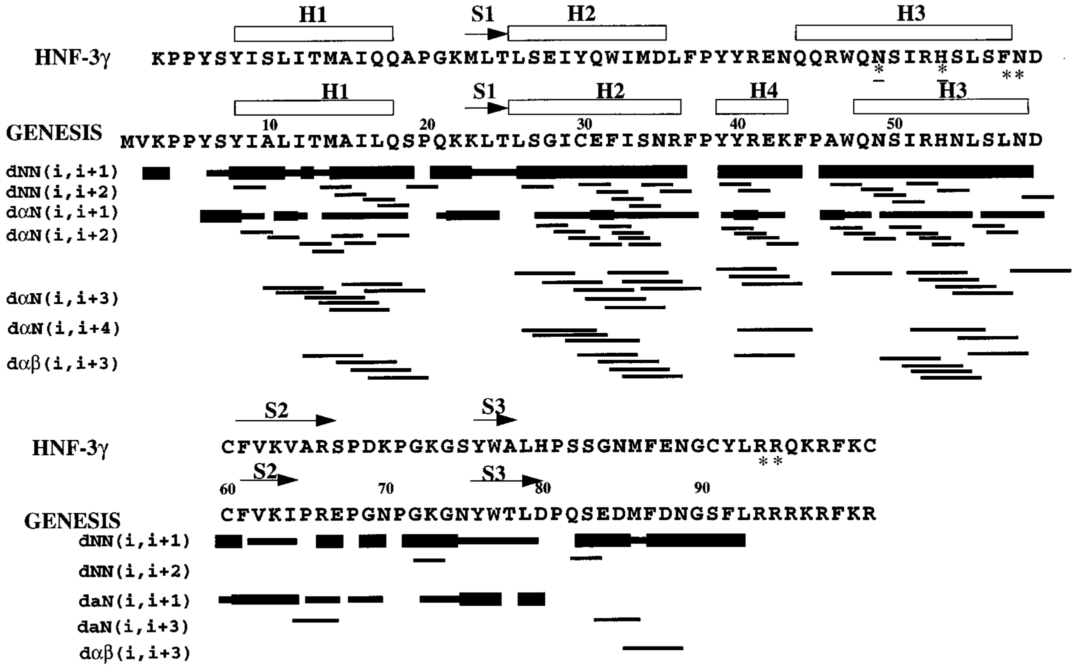

FIGURE 3: Summary of the NOE evidence used to derive the secondary structure of Genesis. The locations of α-helices (H) and β-sheets (S) are marked above and these are compared with the locations of corresponding elements in HNF-3γ, taken from the crystal structure (Clark et al., 1993). Unambiguous NOEs between HN-HN, HN-Hα and Hα-Hβ are marked as $d_{NN}(i, i + n)$, $d_{\alpha N}(i, i + n)$ and $d_{\alpha\beta}(i, i + 3)$, respectively, with the relative NOE strength being reflected by the bar thickness. $\underline{*}$ denotes a residue that shows a direct contact with a DNA base and * denotes a water mediated DNA base contact as observed in the crystal structure.

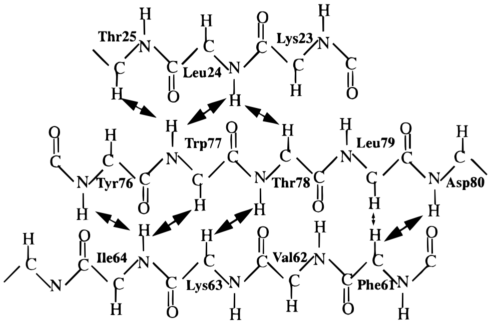

FIGURE 4: Schematic representation of the three-stranded antiparallel β-sheet in Genesis showing the observed NOEs indicated by arrows.

comparison to the other helices and only spans Tyr 39 to Lys 43 as evidenced by $d_{\alpha N}(i, i + 3)$ NOEs for Phe 44, Lys 43, and Glu 42, although a downfield Cα shift is not observed for Phe 44. The observation of this helix, the significance of which will be discussed later, is particularly interesting since this was not observed in the crystal structure of HNF-3γ. A short random coil is seen between H4 and the beginning of H3 at Gln 48 as shown by large downfield Cα shifts and upfield shifts of Hα for Gln 48 and Asn 49. The end of H3 occurs at Asn 58 although there is a discrepancy between Hα chemical shifts, NOEs, and the Cα chemical shift values. On examination of the Cα chemical shift changes, it would appear that H3 extends all the way to Cys 60, while positive $\Delta\delta$ Hα values for Leu 57/Asn 58 might suggest termination at Ser 56. We rationalize that these observations are probably due to a slight deviation in the regular helical structure seen in this region due to some flexibility at the end of the helix. Therefore, given this lack of agreement, we have placed more emphasis on observed characteristic NOEs for α-helices which extend to Asn 58. Following immediately after Cys 60 is the second β-strand (S2), spanning four residues from Phe 61 to Ile 64 and showing NOE cross-peaks to Asp 80, Trp 77, and Tyr 76. These and the aforementioned NOEs show that the final β-strand (S3) occurs between Tyr 76 and Asp 80 and this is consistent with Cα upfield and Hα downfield shifts in all these regions. Similar to other winged helix domains, Genesis also displays two "wings", one comprised of residues between S2 and S3 (Pro 64 → Asn 74) and the other wing of all residues after Pro 81 to the C-terminus. This is shown by very intense signals for these amino acids, particularly for the amides after Pro 81, which are also highly overlapped indicating a highly disordered structure typical of a flexible conformation.

Conservation of H4 in Genesis/DNA Complexes. We have demonstrated the existence of an extra helix (H4) in Genesis when studied in solution as the free protein. However, the observation of this when Genesis is bound to DNA would make for a better comparison with the X-ray structure of the HNF-3γ/DNA complex. With this in mind, we have performed an investigation of a Genesis/DNA complex utilizing a deuterium-labeled protein complexed with a 17 base pair DNA binding site. This DNA binding site contains a 13mer which is the minimum DNA binding sequence for high affinity binding, and an additional four extra base pairs for stabilization of duplex formation (Figure 5c). Although Genesis shows 54% sequence similarity with HNF-3γ and contains homologous residues to those which have been shown to contact DNA in the HNF-3γ/DNA complex, the 13mer core sequence is rather different from the 13mer sequence used in the X-ray crystal study.

The HSQC spectrum of this complex is shown in Figure 6, and as can be seen from this when compared to Figure 1, binding is seen to occur as evidenced by significant chemical shift changes for specific residues, cf. Asp 59: HN, 8.83 ppm; ^{15}N, 123.55 ppm (free); HN, 9.54 ppm; ^{15}N, 124.55 ppm (complex). However, if we now compare the chemical shifts of residues that compose the region where H4 is formed for the free protein and Genesis/DNA complex we see that no significant changes are observed. Furthermore, a com-

```
A.
     *  ** *
  GACTAAGTCAACC
  CTGATTCAGTTGG

B.
  CTTAAAATAACAA
  GAATTTTATTGTT

C.
  GCTTAAAATAACAATAC
  CGAATTTTATTGTTATG
```

FIGURE 5: A comparison of core sequences for high affinity binding of Genesis and HNF-3γ. (A) Binding site used in the X-ray crystal structure of HNF-3γ, $\underline{*}$ denotes a base which has direct contact with a protein residue and * denotes a water mediated protein contact. (B) Minimum binding sequence required for high affinity Genesis binding. Binding site used in our NMR study of a Genesis/DNA complex.

parison of the d_{NN} connectivities in the NOESY spectra show that there is a strong similarity in the NOE patterns, see Figure 7. Therefore, it appears that H4 is still conserved on binding to DNA since we would expect some chemical shifts changes and a loss of NOE correlations were a structural change to occur in this region.

DISCUSSION

In the crystal structure of the HNF-3γ/DNA complex, HNF-3γ folds into a winged helix turn helix motif. Since the DNA binding domain of Genesis has 54% sequence homology with HNF-3γ, it is also expected to fold in a similar manner. However, like many other family members, it contains several glycine and proline residues within regions that in HNF-3γ are shown to be α-helix or β-sheet. Therefore, the effect of these residues on the secondary structure of Genesis is unpredictable.

From data taken on a deuterium labeled sample, a regular ^{1}H NOESY-HSQC, and an HNCA, we were able to determine the secondary structural elements of Genesis. Figure 3 shows the proposed secondary structure and a summary of the NOE constraints used in determining this. Also shown for comparison is HNF-3γ. As can be seen from this, the overall folding of Genesis and HNF-3γ are similar in that both proteins display at least three α-helices, three β-strands, and two flexible regions (wings).

On closer examination of the secondary structure of both HNF-3γ and Genesis, we see that for H1, H2, and S1, there is virtually no difference in the positions or the length of these motifs except in H2, which is one amino acid longer in Genesis. The major differences occur, however, for structural motifs that form the primary DNA recognition region, i.e., helix 3. H3 in HNF-3γ binds in the major groove of DNA and makes two base-specific contacts (residues shown in Figure 3). It is 14 amino acids in length and starts at Gln 44 and extends to Phe 57 (note: our numbering scheme for HNF-3γ differs from that in the published crystal structure for better comparison with structural elements in Genesis). In contrast, H3 in Genesis is slightly smaller and right shifted by four amino acids in that it is only 11 amino acids in length and begins at Gln 48. Furthermore, Genesis displays an additional helix (H4) between H2 and H3 where HNF-3γ shows only coil. These observations are significant for DNA binding since H3 provides the principal DNA contacts on binding to the major groove.

In our secondary structure, three β-strands are also observed, the same number as in HNF-3γ. S1 is in the same position as in HNF-3γ, and a short strand (S2) is observed between Phe 60 and Ile 63, which is three amino acids shorter than the corresponding S2 in HNF-3γ. This short strand is possibly due to the presence of a proline residue within the sequence at position 65. Interestingly, a longer S3 is observed in Genesis than in HNF-3γ and due to the shortening of S2, a longer W1 between S2 and S3 is also observed, again probably due to the different positions of Pro and Gly. On the basis of cross-strand HN-HN and HN-sidechain NOEs, the order of S1, S2, and S3 is similar to those observed in HNF-3γ, being antiparallel with a S1-S3-S2 arrangement.

HFH proteins have highly conserved DNA binding regions, and this is especially true among the amino acids

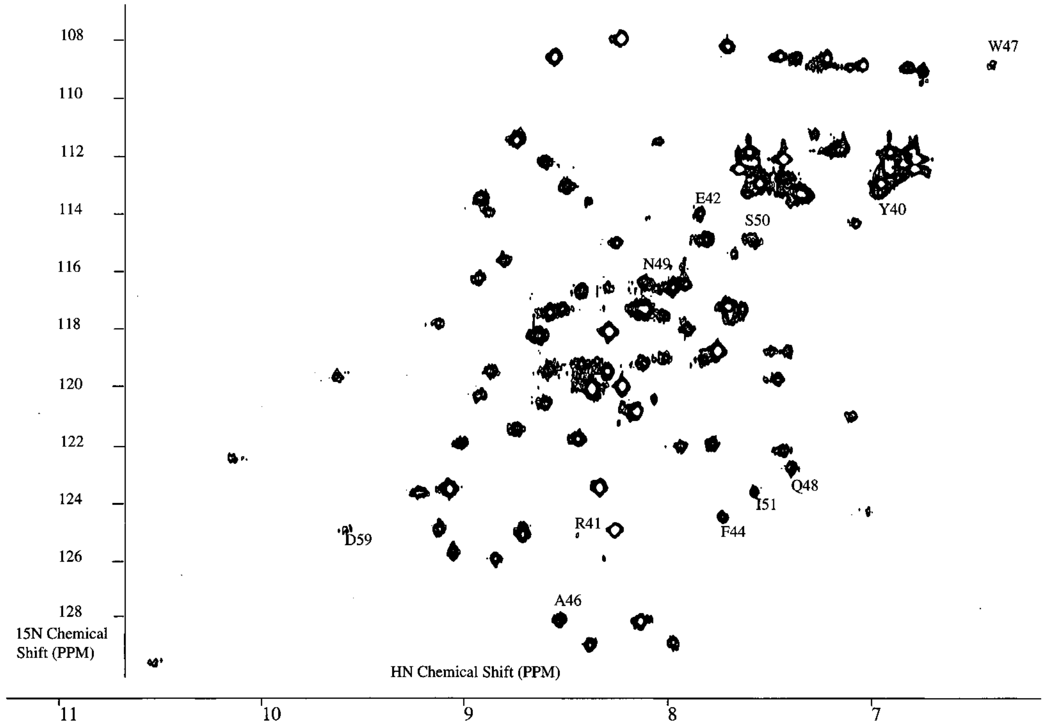

FIGURE 6: ^{1}H-^{15}N HSQC spectrum of Genesis complexed to DNA. The 17mer DNA sequence used is shown in Figure 5C.

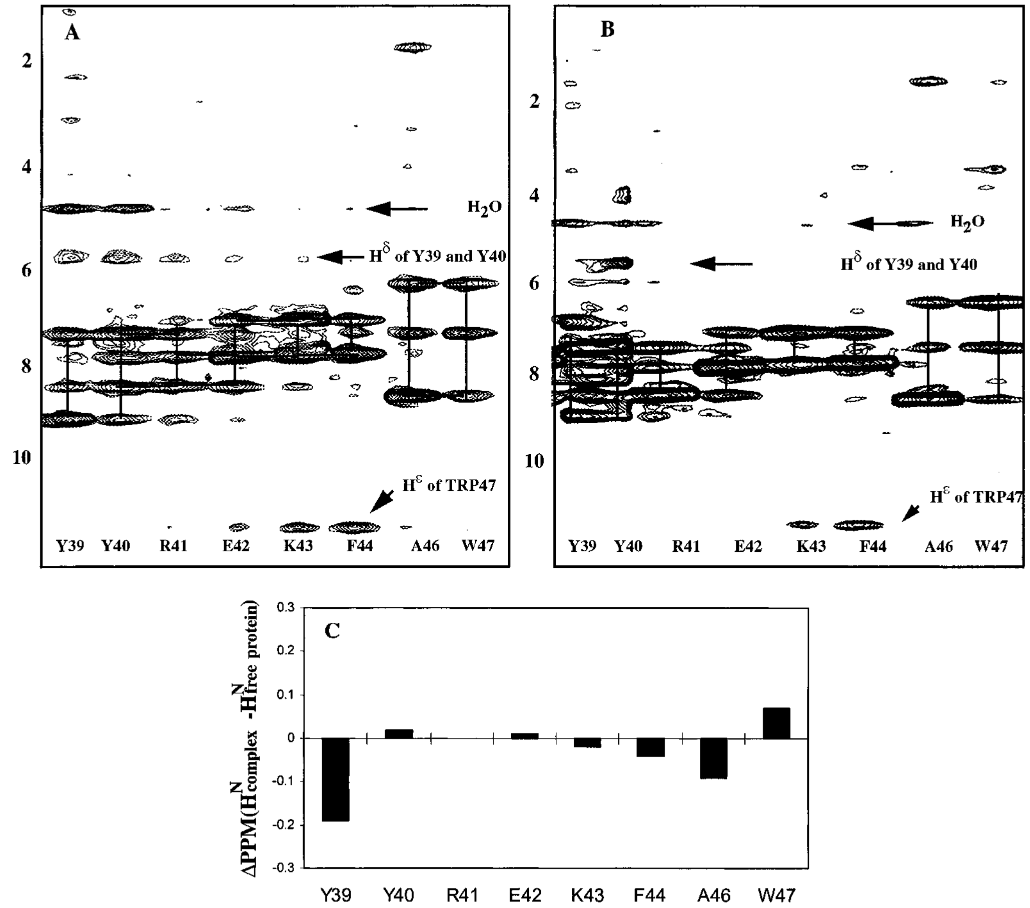

FIGURE 7: Stripcharts showing the backbone (d_{NN}) NOE connectivities of the amino acid residues from Tyr 39 to Trp 47 for Genesis free in solution (A) and bound with its cognate DNA sequence (B). This sequence represents the region of the protein which forms H4 in the free protein and as can be seen from a comparison of these two charts, very little perturbation of the ^{1}H chemical shifts occurs on complexation of DNA. (C) Shows $\Delta\delta$ ppm ($\Delta\delta$ ppm Genesis/DNA $-$ Genesis$_{free}$) values between Genesis free in solution and bound to DNA in the region of H4.

of the recognition helix H3, which is found to be almost invariable. However, despite this, different family members have their own unique DNA binding specificity. This was shown using chimeric proteins, where Overdier et al. (1994) demonstrated that a 20 amino acid sequence adjacent to the recognition helix influences the DNA binding specificity in different HNF-3 family members. Here, replacement of a 20 amino acid region in HNF-3β with corresponding residues from Genesis was sufficient to cause Genesis specific binding. This 20 amino acid sequence stretches from the C-terminus of helix 2 to the N-terminus of helix 3, including all intervening residues, and is one of the most diversified sequences within the different family members. However, due to conserved residues at distinct positions within this sequence, these proteins can be divided into different subfamilies. Therefore, a hypothesis was proposed that this 20 amino acid sequence regulates the presentation of helix 3, thereby subtly altering the binding specificity. Our NMR studies support such a hypothesis; as the major differences observed between the crystal structure of HNF-3γ and our NMR data occur in this region, where we demonstrate that in the middle of the 20 amino acid sequence, an extra small helix (H4) is formed. In this 20 amino acid sequence in Genesis, there are two regions that diverge significantly from the sequence in HNF-3γ, these are between Ser 34$-$Arg 36 and Glu 42$-$Ala 46. Interestingly, however, we actually observe H4 in a conserved sequence (FPYYR) between these two diverse sequences, a sequence that is also relatively well conserved between different subgroups. This sequence in HNF-3γ, although not showing a distinct helix, shows some twist, and so it is possible that this sequence is the point where the presentation of H3 to DNA is mediated based on its propensity to form an α-helix. This might depend on adjacent residues as to whether a helix is formed and how long this actually is, but given the diversity of these flanking sequences, such a mechanism could lead to many ways of tuning the binding specificity of H3.

In the report of the binding of different family members where the 20 amino acid sequence was proposed to influence the specificity (Overdier et al., 1994), no attempt was made to determine the least number of amino acids that could be inserted in chimeric proteins to still retain the binding characteristics of the inserted proteins amino acid sequence. Significantly, Pierrou et. al. (1994), in their study of human forkhead proteins, FREAC-1-7, observed that preferential binding of a cytosine at a specific position in a core sequence was encoded by a region in the central part of the forkhead domain. Again, using chimeric proteins, they rationalized that differences in binding specificity could be attributed to differences in the primary structure occurring in the loop between H2 and H3, and in the first three amino acids of H3 i.e., F-DNKQG in FREAC-3 and Y-EKFPA in FREAC-4. This is precisely the region where we observe the formation of H4 and a shortening of H3 for the exact same sequence in Genesis as in FREAC-4, i.e., YYREKFPA. Therefore, it is possible that a shorter substituted amino acid sequence would still show the same effects, and this would be consistent with what we and Pierrou et al. observe.

The formation of H4, or lack thereof in HNF-3γ, could be due to the differences in the free (Genesis) and DNA-bound (HNF-3γ) protein conformation and not necessarily any unique property associated with binding specificity. However, we have also performed investigations of Genesis bound with its cognate DNA sequence, and after binding DNA, the amide ^{15}N and ^{1}H chemical shifts in the H4 region are largely unperturbed (Figures 1, 6, and 7). This suggests that H4 is also present in the DNA complex, and therefore, the differences observed are not merely due to conformation changes on binding to DNA.

An interesting observation which may be significant for overall binding of HFH proteins is that residues from 82 onward show strong NOEs and slightly higher upfield $\Delta\delta$ chemical shift changes than would be expected simply from random coil values. However, due to the lack of a significant number of characteristic NOEs for α-helices such as $d_{NN}(i, i+2)$, $d_{\alpha N}(i, i+3)$, and $d_{\beta N}(i, i+3)$ we feel that a stable helix is not formed, rather, what may be formed is some type of preferred coiled conformation. It is known from the crystal structure that residues in this flexible wing contact DNA in the minor groove, and so it is possible that such a preferred orientation acts as a secondary recognition motif in the binding process.

In summary, using NMR, we have determined the secondary structure of the DNA binding domain of the protein Genesis. This is very similar to that observed in the crystal structure of a homologue, HNF-3γ, except in the recognition region where Genesis shows two smaller helices instead of one large helix. This may account for the different DNA binding specificities.

ACKNOWLEDGMENT

I.M. acknowledges helpful discussion and assistance from Dr. B. Volkman and Dr. F. Abildgaard, NMRFAM, and we would like to thank Dr. P. W. Gettins, UIC, for critical review of the manuscript. HFH clones and the DNA binding site were kindly provided by Dr. R. H. Costa and Dr. D. Overdier, and finally, we would also like to thank Dr. K. L. Clark and Dr. S. K. Burley for making available the structural coordinates of the HNF-3γ/DNA complex prior to deposit in the protein databank and T. Shiyanova for running a gel shift assay.

SUPPORTING INFORMATION AVAILABLE

This additional information consists of a table of NMR resonance assignments for backbone atoms, HN, Hα, ^{15}N, Cα, and Cβ (3 pages). Ordering information is given on any current masthead page.

REFERENCES

Ang, S. L., & Rossant, J. (1994) *Cell, 78*, 561−574.

Clark, K. L., Halay, E. D., Lai, E., & Burley, S. K. (1993) *Nature 364*, 412−420.

Clevidence, D. E., Overdier, D. G., Tao, W., Qian, X., Pani, L., Lai, E., & Costa R. H. (1993) *Proc. Natl. Acad. Sci. U.S.A. 90*, 3948−3952.

Costa, R. H., & Grayson, D. R. (1991) *Nucleic Acids Res. 19*, 4139−4145.

Costa, R. H., Grayson, D. R., & Darnell, J. E., Jr. (1989) *Mol. Cell. Biol. 9*, 1415−1425.

DeMassy, B., Rocco, V., & Nicolas, A. (1995) *EMBO J. 14*, 4589−4598.

Gardner, K. H., Rosen, M. K., & Kay, L. E. (1997) *Biochemistry 36*, 1389−1401.

Gehring, W. J., & Hiromi, Y. (1986) *Annu. Rev. Genet. 20*, 147−173.

Grzesiek, S., & Bax, A. (1992) *J. Am. Chem. Soc. 114*, 6291−6293.

Farmer, B. T., & Venters, R. A. (1996) *J. Biomol. NMR. 7*, 59−71.

Fesik, S. W., & Zuiderweg, E. R. (1990) *Q. Rev. Biophys. 23*, 97−131.

Harrison, S. C., & Aggarwal, A. K. (1990) *Annu. Rev. Biochem. 59*, 933−969.

Johnson, P. F., & McKnight, S. L. (1989) *Annu. Rev. Biochem. 58*, 799−839.

Kaufmann, E., & Knöchel, W. (1996) *Mech. Dev. 57*, 3−20.

Kay, L. E., Ikura, M., Tschudin, R., & Bax, A. (1990) *J. Magn. Reson. 89*, 496−514.

Kornberg, T. B. (1993) *J. Biol. Chem. 268*, 26813−26816.

Lai, E., Prezioso, V. R., Smith, E., Litvin, O., Costa, R. H., & Darnell, J. E., Jr. (1990) *Genes Dev. 4*, 1427−1436.

Landschulz, W. H., Johnson, P. F., & McKnight, S. L. (1988) *Science 240*, 1759−1764.

Liu, E., DiPersio, C. M., & Zaret, K. S. (1991) *Mol. Cell. Biol. 11*, 773−784.

Miller, J., McLachlan, A. D., & Klug, A. (1985) *EMBO J. 4*, 1609−1614.

Mitchell, P. J., & Tjian, R. (1989) *Science 245*, 371−378.

Muhandiram, D. R. & Kay, L. E. (1994) *J. Magn. Reson., Ser. B 103*, 203−214.

Murre, C. McCaw, P. S., & Baltimore, D. (1989) *Cell 56*, 777−783.

Overdier, D. G., Porcella, A., & Costa R. H. (1994) *Mol. Cell. Biol. 14*, 2755−2766.

Pabo, C. O., & Sauer, R. T. (1992) *Annu. Rev. Biochem. 61*, 1053−1095.

Palmer, A. G., Cavanagh, J., Wright. P. E., & Rance, M. (1991) *J. Magn. Reson. 93*, 203−216.

Pierrou, S., Hellqvist, M., Samuelsson, L., Enerbäck, S., & Carlsson, P. (1994) *EMBO J. 13*, 5002−5012.

Smith, B. O., Ito, Y., Raine, A., Teichmann, S., Ben-Tovim, L., Nietlispach, D., Broadhurst, R. W., Terada, T., Kelly, M., Oschkinat, H., Shibata, T., Yokoyama, S., & Laue, E. D. (1996) *J. Biomol. NMR., 8*, 360−368.

Sutton, J., Costa, R., Klug, M., Field, L., Xu. D., Largaespada, D. A., Fletcher, C. F., Jenkins, N. A., Copeland, N. G., Klemsz M. & Hromas, R. (1996) *J. Biol. Chem. 271*, 23126−23133.

Tabor, S., & Richardson, C. C. (1987) *Proc. Natl. Acad. Sci. U.S.A., 84*, 4767−4771.

Venters, R. A., Huang, C. C., Farmer, B. T., II, Trolard, R., Spicer, L. D., & Fierke, C. A. (1995) *J. Biomol. NMR. 5*, 339−344.

Weigel, D., & Jäckle, H. (1990) *Cell 63*, 455−456.

Weinstein, D. C., Ruiz i Altaba, A., Chen, W. S., Hoodless, P., Prezioso, V. R., Jessell, R. M., & Darnell, J. E., Jr. (1994) *Cell 78*, 575−588.

Wishart D. S., & Sykes, B. D. (1994a) *Methods. Enzymol. 239*, 363−392.

Wishart D. S., & Sykes, B. D. (1994b) *J. Biomol. NMR. 4*, 171−180.

Wittekind, M., & Mueller, L. (1993) *J. Magn. Reson., Ser. B 101*, 201−205.

Wuthrich, K. (1986) *NMR of Proteins & Nucleic Acids*, Wiley Interscience Publication, New York.

Zenvirth, D., Arbel, T., Sherman, A., Goldway, M., Klein, S., & Simchen, G. (1992) *EMBO J. 11*, 3441−3447.

BI971514M