COMMUNICATION

Structural Changes in the Region Directly Adjacent to the DNA-binding Helix Highlight a Possible Mechanism to Explain the Observed Changes in the Sequence-specific Binding of Winged Helix Proteins

Ian Marsden, Changwen Jin and Xiubei Liao*

Department of Biochemistry and Molecular Biology, College of Medicine, 1819 West Polk Street, University of Illinois at Chicago, Chicago, IL 60612 USA

The hepatocyte nuclear factor 3 (HNF-3)/fork head (fkh) family contains a large number of transcription factors and folds into a winged helix motif. Despite having almost invariable amino acid sequences in their principal DNA-binding helices, HNF-3/fkh proteins show a wide diversity of sequence-specific binding. Previous studies of chimeric HNF-3/fkh proteins demonstrated that the binding specificity is primarily influenced by a region directly adjacent to the binding helix. We report our findings of an NMR structural study performed on an HNF-3/fkh family member (Genesis, formerly HFH-2) and compare it to that of another family member (HNF-3γ) complexed to DNA and determined by X-ray crystallography. It is found that in comparison to HNF-3γ, Genesis contains an extra small helix directly prior to the N terminus of the primary DNA contact helix. Due to the insertion of this helix, a shorter and slightly re-positioned primary DNA contact helix is observed, which we believe leads to the DNA-binding specificity differences among family members.

The winged helix DNA-binding motif was initially identified in the products of the Drosophila homeotic gene forkhead and hepatocyte nuclear factor 3 (Weigel & Jäckle, 1990; Lai et al., 1990). Since these initial discoveries, many proteins containing the HNF-3γ/fkh homologous winged helix motif have been identified in a variety of organisms ranging from yeast to human and their function is linked to gene regulation during embryogenesis, tumorigenesis and in differentiated cell states (reviewed by Kaufmann & Knöckel, 1996). A structure determination of one of the HNF-3γ/fkh homologous proteins, an HNF-3γ/DNA complex, displays three α-helices, three β-strands and two flexible loops, or ‘wings’ in the C-terminal region (Clark et al., 1993). The protein binds as a monomer where helix 3 (H3) is the principal DNA recognition site and makes contacts to the DNA site in the major groove. In addition, the flexible C-terminal loops interact with the phosphate backbone in the minor groove. Given the structure of the HNF-3γ/DNA complex, an interesting observation is that although the proteins in this family are highly conserved, especially the amino acid sequences in the DNA recognition helix, these proteins demonstrate divergent DNA-binding specificity. On the basis of sequence swap experiments, Overdier et al. (1994) showed that a stretch of divergent amino acid sequence prior to the DNA recognition helix regulates DNA-binding specificity in these family members. This result provides an important insight into protein-DNA recognition in HNF-3/fkh homologous transcription factors and one that deserves systematic evaluation.

One of these proteins, the product of Genesis, formerly HFH-2 (Clevidence et al., 1993), has been shown to be specifically expressed in embryonic stem cells or their malignant equivalent and co-transfection experiments revealed also that Genesis is a transcriptional repressor (Sutton et al., 1996). Although the amino acid sequence in helix 3 of the
DNA-binding domain of Genesis is almost identical with that of HNF-3γ, the two proteins show different DNA-binding specificity. By using NMR techniques, we demonstrated that the sequence prior to helix 3 in Genesis forms an extra short helix (helix 4) in both the Genesis/DNA complex and the free protein, instead of the random coil observed in HNF-3γ (Marsden et al., 1997).

The secondary structure of Genesis can be readily obtained by qualitatively evaluating short-range NOE patterns and chemical shift values; however, a tertiary structure determination requires a more quantitative interpretation of NOE intensities. In this respect, studying HNF-3/fkh homologues by NMR presents many difficulties, one of which is highly uneven signal intensity and signal overlap due to a broad range of mobility differences in these proteins, which contain two long flexible sequences. Aided by partial deuteration of the protein it is possible to overcome some of these difficulties, and we present the solution structure of Genesis as determined by NMR spectroscopy.

**NMR spectroscopy**

The expression and purification of stable isotope-enriched Genesis has been described (Marsden et al., 1997). NMR spectra of the free protein were recorded on either a Varian Unity Plus 500 or Bruker DMX 500 NMR spectrometer at 290 K and pH 6.5, except one duplicate 15N-edited NOESY-HSQC, which was acquired at 296 K. The NMR data were processed and analyzed using Triad 6.2 software (Tripos, Inc., St. Louis, MO).

The backbone and side-chain assignments were obtained according to standard procedures and have been published (Marsden et al., 1997). NOE constraints were obtained from the following experiments: 3D 15N-edited NOESY-HSQC in H2O (Fesik & Zuiderweg, 1990), on both proteins grown in 100% H2O and 98% 2H2O with mixing times of 140 and 180 ms respectively. A 3D 1H-13C-correlated NOESY-HSQC in 2H2O (mixing time 140 ms) and a homonuclear NOESY experiment in 2H2O were also acquired (mixing time 140 ms). Sensitivity-enhanced gradient pulse sequences were employed for all experiments in which magnetization was detected on the amide HN (Muhandiram & Kay, 1994; Palmer et al., 1991).

**Collection of conformational constraints for structure calculations**

The interproton distance restraints were derived from the analysis of four different NOE spectra: (i) and (ii) NOEs involving NH protons were extracted from 3D 15N NOESY-HSQC spectra measured in H2O of both fully protonated and partially deuterated proteins; (iii) NOEs between aliphatic protons were taken from a 3D 13C NOESY-HSQC spectrum measured in 2H2O; (iv) data involving aromatic protons were determined from a 2D NOESY spectrum in 2H2O solution.

Distance constraints were estimated using the criteria of small, medium and large NOEs after taking into account different cross-peak intensities in the different spectra used, i.e. normalizing cross-peak intensities with respect to the overall sensitivity of the spectrum due to the use of both protonated and deuterated proteins. Therefore, the distances that were used for small, medium and large Hα-HN correlations for the deuterated protein were 2.80 Å, 4.50 Å and 6.00 Å, respectively. For the protonated proteins, upper limits for backbone-backbone NOE correlations were set up to 3.00 Å, 4.00 Å and 5.00 Å, while 3.5 Å, 4.5 Å and 5.5 Å were used for backbone–side-chain and side-chain–side-chain NOEs. These values were also the base values used to make pseudo-atom corrections for moieties where stereospecific assignments were not available.

Structure calculations were performed with the program DIANA (Güntert et al., 1991). As indicated previously (Rosen et al., 1996), the deuterated Genesis protein in our study has selective side-chain proton enrichment at known positions (Marsden et al., 1997). Thus, initial structures were calculated using only data taken from the deuterated protein. These initial structures were then examined to make further NOE assignments, which were then incorporated into successive calculations. The input for these calculations used a total of 1195 upper distance constraints distributed as follows, 230 intraresidual, 311 sequential, 340 medium and 314 long-range correlations: Figure 1 shows the sequence distribution. Since the deuterated protein is only partially protonated at specific sites and due to increased NOE signal intensities and improved signal uniformity, the cross-peak intensities between two protons in the NOESY spectrum of deuterated Genesis are much better correlated to the actual distance between the two protons. Thus we were able to include 47 lower limit constraints to improve the structure quality in the β-strand region. These lower limits were included only after a majority of the upper distance constraints were incorporated into the calculation.

Due to fast T2 relaxation in the protonated molecule at 290 K we were unable to gain dihedral angle data between Hα and Hβ through the use of a (HNHA) experiment, as few signals were observed. However, since Cα chemical shift index values can be directly correlated to the secondary structure motif we incorporated 61 angle constraints on the basis of the chemical shift index (Wishart & Sykes, 1994a,b). The φ angle was constrained between −70° and −140° when Cα -obs minus Cα-random is smaller than −1 and the angle was constrained between −40° and −80° when Cα -obs minus Cα-random is larger than 1. These criteria are commonly used to define secondary structure and these angle constraints were also incorporated at the latter stages of the calculations. Incorporation of both lower limits and angle constraints did not cause any undue violations of the upper limit.
constraints we had used initially and improved the structure quality for β-sheet regions slightly.

A large number of constraints were observed in the α-helical regions, while in contrast, few medium or long-range constraints were found for the C-terminal region, indicating that this is highly flexible in solution and is not involved in interactions with the protein core. The N-terminal region also is not well defined, although some long-range correlations were observed, showing that it is at least constrained to a certain extent to be in proximity to the end of H2 and the start of H4.

The final ensemble of structures represented by the 20 best DIANA conformers (Figure 2) from an input of 100 initial structures is analyzed in Table 1. The coordinates have been deposited in the Brookhaven Protein Databank (Identification number, 2HFH). When the 20 structures are fit using atoms in well-structured regions, i.e. residues 8 to 64, 75 to 80, the RMSD of backbone atoms is found to be 0.66 Å, while the RMSD of heavy atoms is 1.23 Å. This is in sharp contrast to the full protein, which has RMSDs for backbone and heavy atoms of 5.46 Å and 5.93 Å, respectively. The local RMSD values are shown in Figure 1. The only NOE violation between 0.4 and 0.5 Å (0.47 Å) is in structure 20. The stereochemical quality of the protein was also assessed with the program PROCHECK (Laskowski et al., 1993) on the best 20 DIANA structures. These results indicate using 93 residues for which resonance assignments were made, 64.9% of residues are in favored regions while no residue was found to be contained within disallowed regions (Table 1). Due to the lack of NOE constraints, many residues in the C terminus of the protein are in the generously allowed region, which contains 10.4% of the total residues. Further data also show that the intervening region in the protein, i.e. wing 1, is slightly more disordered than the helices and sheets and that wing 2 is highly disordered (Figures 1 and 2). Previously we demonstrated that Genesis recognizes a prebent DNA site (Bravieri et al., 1997). Therefore, it seems likely that this flexibility may be the structural basis for this property.

Figure 2. A stereo diagram showing the superposition of the 20 best calculated DIANA structures of Genesis.
The solution conformation of Genesis

The schematic representation of the solution structure of Genesis is shown in Figure 3A and overlaid and compared to the crystal structure of HNF-3γ in Figure 3C. It is comprised of four α-helices (three long and one short) and one β-sheet, which is formed of three antiparallel β-strands. All four α-helices appear in the N-terminal part of the protein and in addition to the regular secondary structural elements there are 2 flexible loops or “wings” in the C-terminal region, hence the term winged helix. The α-helices were identified on the basis of characteristic NOE patterns (Wüthrich, 1986), the chemical shift index (Wishart & Sykes, 1994a,b) and hydrogen bonding patterns between the carbonyl oxygen atom of residue i and the amide proton of residue i+4 where this was observed in a majority of the 20 lowest energy structures calculated from DIANA (Güntert et al., 1991). Therefore, the helices were identified as occurring at residues Tyr8 to Gln18 (H1), Leu26 to Arg36 (H2), Pro38 to Phe44 (H4) and Trp47 to Leu57 (H3). Notice that H4 precedes H3 in the amino acid sequence, which for consistency we have labeled based on considerations of the HNF-3γ crystal structure, which shows only three helices (Figure 3B), H3 being the principal DNA-binding helix. In addition, the protein shows a three-stranded β-sheet at residues Lys23 to Thr25 (S1), Phe61 to Ile64 (S2) and Tyr76 to Asp80 (S3), where S1 is twisted slightly out of planarity with S2 and S3. The positions of these secondary structural elements coincide quite closely with our previously published secondary structure (Marsden et al.,

<table>
<thead>
<tr>
<th>Table 1. Statistics of 20 DIANA structures of GENESIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backbone</td>
</tr>
<tr>
<td>RMSD (Å) (1-93)</td>
</tr>
<tr>
<td>RMSD (Å) (8-64, 75-80)</td>
</tr>
<tr>
<td>NOE violations</td>
</tr>
</tbody>
</table>

Ramachandran analysis for non-Gly and non-Pro residues (1-93)
- Residues in most favored regions (%) | 64.9
- Residues in additional allowed regions (%) | 24.7
- Residues in generously allowed regions (%) | 10.4
- Residues in disallowed regions (%) | 0.0

Figure 3. A, A ribbon diagram showing the NMR-derived tertiary structure of Genesis. B, A ribbon diagram showing the X-ray crystal structure of HNF-3γ (Clark et al., 1993). C, A stereo diagram of the backbone Cα atom coordinates of the core helices of Genesis (green) overlaid with those of HNF-3γ (red). Diagrams were produced with MOLMOL, a molecular graphics program for displaying and analyzing the 3D structures of biological molecules (Koradi et al., 1996).
1997), determined solely by the chemical shift index method and examination of characteristic NOEs; however, there are some slight differences. In comparison with the previous data, H1 also begins at Tyr8 and ends at Gln18; H2 has the same length, beginning at Leu26 and ending at Arg36; H4 meanwhile is slightly longer, again beginning at Pro38 but ending one residue later at Phe44 and finally H3 is again displaced compared to H3 in our previous report, this time by two residues, in that it begins at Trp47 and ends at Leu57 cf. Asn49 to Asp59.

The three large α-helices display a triangular arrangement with the fourth smaller helix essentially making up an extended apex of this triangle, while the β-sheet forms a planar base of the protein above which the helices are located. To either side of the β-sheet the two wings extend outward with wing 1 being better defined than wing 2, which is completely flexible in solution, as evidenced by intense signals in triple resonance experiments and lack of long-range NOEs. The core of the protein involves 14 hydrophobic amino acid residues in both the α-helices and β-sheet, since the core of the four-helix bundle is located just above the plane of the β-sheet. The residues in question are Tyr8, Leu11, Ile12, Ile16, Ile29, Phe32, Ile33, Tyr40, Trp47, Ile51, Leu55, Phe61 and Trp77. All of these amino acids are either invariant (Ile29, Trp47, Leu55, Phe61 and Trp77) or highly conserved in the winged helix motif of the mammalian HNF-3/fkh transcription factor family. This suggests that most, if not all, members of this family should adopt a global fold similar to that of HNF-3γ with at least three α-helices and a β-sheet.

Comparison of the NMR solution structure of Genesis and X-ray crystal structure of HNF 3γ

The only structural study of a HNF-3γ/fkh family member to date is a crystal structure study of HNF-3γ complexed with DNA, determined by X-ray crystallography at 2.5 Å resolution (Clark et al., 1993). Therefore, since the DNA-binding domain of Genesis has 54% sequence homology with HNF-3γ, it is interesting to compare the NMR structure with this crystal structure to try to gain some insight into the structural basis for the different binding specificities observed between Genesis and HNF-3γ. Like Genesis, HNF-3γ contains both α-helices and a three-stranded β-sheet (Figure 3B), although Genesis displays an extra small helix (H4) compared to HNF-3γ (Figure 3A). Given this, however, a mean structure derived from an ensemble of the 20 structures with the lowest target function from DIANA can be superimposed on the crystal structure coordinates to give an RMSD for the well-structured regions, i.e. residues 8 to 64, of 0.930 Å for all backbone atoms, indicating that the two structures are very similar.

This similarity is seen if we compare the positions and lengths of the various secondary structural elements observed in both Genesis and HNF-3γ (Figure 1, bottom). The major difference between the two structures occurs in the region of helix 3, where in Genesis H3 is smaller than in HNF-3γ and adjacent to this we observe another small helix (H4). The longer helix (H3), which putatively binds DNA in the major groove, starts at Trp47 and ends at Leu57, the same residue position as in HNF-3γ, although by starting at position 47 this makes the helix some three residues shorter in Genesis. This discrepancy may be accounted for by the presence of H4, which is directly adjacent to H3 and which spans seven residues between Pro38 and Phe44. On examination of the superposition of H1 to H3 (Figure 3C), the average distance between pairs of Cα atoms in each helix is: H1 1.13 (±1.40) Å, H2 0.97 (±0.36) Å and for H3 1.80 (±1.77) Å, showing that H3 has the largest difference in superpositioning between Genesis and HNF-3γ (Figure 3C). This is particularly true for the C-terminal section of this helix, where the greatest divergences occur; that is, for residues 57 and 58 the pairwise difference between the Cα atoms of the two proteins is 3.57 and 3.13 Å, respectively. These observations may be significant for DNA binding, since the C terminus of helix 3 makes multiple DNA contacts in HNF-3γ and likely in a Genesis-DNA complex (unpublished results).

Implications of structural differences between Genesis and HNF-3γ on DNA binding specificity

HNF-3γ/fkh homologous transcription factors have highly conserved DNA-binding regions and this is especially true for the amino acid residues of the recognition helix H3, which is almost invariable among many family members. However, despite this, the binding specificity and core DNA-binding sequences are significantly different. An understanding of the basis for sequence-specific binding is therefore necessary for a complete comprehension of how this family of proteins functions. Our observation of a fourth helix (H4) in Genesis compared with HNF-3γ, see Figure 3, may have significant implications for DNA sequence recognition and binding specificity, especially since this helix is adjacent to the principal helix (H3) that binds to DNA in the major groove.

A previous report by Overdier et al. (1994) found that the DNA-binding specificity of the HNF-3/fkh domain is influenced by amino acid residues outside the recognition helix, H3. This was demonstrated through the use of chimeric proteins, whereby replacement of various stretches of amino acid regions in HNF-3γ with corresponding residues from another homologue, HFH-1, was sufficient to cause HFH-1-specific binding in some cases. Through the use of a series of different chimeras, it was found that a region consisting of 20 amino acid residues adjacent to the recognition helix was the primary controlling element affecting DNA-binding specificity. These results demonstrate that this region is capable of causing a recognition helix from one protein to bind
DNA sites that are specific for another protein, possibly by alteration of the base-specific contacts or the presentation of this helix such that alternative residues participate in binding. This 20 residue sequence stretches from the C-terminal half of helix 2 to the first four residues of H3, and includes all intervening residues. Interestingly, this is the region in which we observe helix 4 and the region in which the largest sequence variability among family members occurs.

Numerous chimeric proteins were made to investigate the binding properties of HFH proteins, although a comprehensive series was not used to determine the least number of amino acid residues that influence binding. However, in a similar study by Pierrou et al. (1994) of human fkh proteins, FREAC 1 to 7, it was also observed that preferential binding of a cytosine base at a core sequence was encoded by a region in the central part of the domain and this region was narrowed down to the loop between H2 and H3, and the first three residues of helix 3. Again, this is precisely the region where we observe the major structural differences between Genesis and HNF-3γ, with formation of helix 4 and a concomitant shortening and repositioning of helix 3.

These NMR studies would seem to support the hypothesis of Overdier et al. (1994), that the amino acid sequence adjacent to the recognition helix is responsible for determining the DNA-binding specificity, in that we observe an extra helix in this region when compared with the structure of HNF-3γ from X-ray crystallography. This is the only major difference between our NMR structure of Genesis and that of HNF-3γ, although the actual extent of this region is somewhat undefined, given that an investigation of a comprehensive set of chimeric proteins has not been studied. Indeed, the size of this region may vary among the different family members but, given our results, it might be related to the formation of a helix in this region that is structurally capable of regulating the binding of helix 3 in some manner. In the crystal structure, this region is twisted into a quasi-helical loop and, depending on the amino acid residues in this area, an α-helix may form to a greater or lesser degree thereby regulating DNA binding. As mentioned, a comparison of the superposition of helix 3 in both proteins shows the largest divergence between pairs of Cα atoms especially toward the C terminus, supporting the idea that the recognition helix is displaced through the formation of helix 4.

Further evidence in favor of this hypothesis is that the 20 amino acid residue region adjacent to H3 in the HNF-3/fkh family shows high sequence variability, except for the middle section from Phe37 to Arg41 in Genesis. These five residues are highly preserved in HNF-3/fkh winged helix homologues, with only conservative substitutions among each protein. However, the sequence prior to this, which forms the C terminus of H2, and after the N terminus of H3, both show wide sequence variations. Since HNF-3γ shows a quasi-helical loop and Genesis shows a helix spanning the above five amino acid residues, it is possible that this section has a propensity to form an α-helix where the residues flanking this section determine its length and act to cause interactions within this whole section. Therefore, such a mechanism might be capable of mediating both the position of H3 and the possible interactions between H2 and H3.

The NMR structural study of Genesis was performed on the free protein, whereas the crystal structure study was of HNF-3γ bound to its cognate DNA site; therefore, it is possible that the differences we observe between the two proteins are solely due to this. However, results from HN and Cα chemical shift data on a Genesis/DNA complex display minimal chemical shift changes in this region compared to the free protein (unpublished results), which would seem to suggest that helix 4 is still formed when bound to DNA thereby disfavoring this possibility.

Acknowledgments

This work was funded by grant GM 52034 from the National Institutes of Health and a grant (95010100) from the American Heart Association. X.L. is also the recipient of a JFA from the American Cancer Society. We thank Dr Y. Chen and Dr M. Rance for use of their NMR to acquire the data for this research. We thank Dr R. Costa for the gift of the Genesis gene and fruitful discussion. We thank Dr P. Gettins for reviewing the manuscript and critical comments. This study also 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.

References


*Edited by P. E. Wright*

(Received 13 November 1997; received in revised form 30 January 1998; accepted 2 February 1998)