Homology Model for Oncostatin M Based on NMR Structural Data

Douglas Kitchen,[‡,§] Ross C. Hoffman,[||,⊥] Franklin J. Moy,[‡] and Robert Powers*[,‡]

Immunex Corporation, 51 University Street, Seattle, Washington 98101, and Department of Structural Biology, Wyeth-Ayerst Research, Pearl River, New York 10965

Received November 6, 1997; Revised Manuscript Received March 5, 1998

ABSTRACT: Oncostatin M (OM) is a member of the cytokine family which regulates the proliferation and differentiation of a variety of cell types and includes interleukin-6 (IL-6), leukemia inhibitory factor (LIF), and granulocyte-colony stimulating factor (G-CSF). This family of proteins adopts a four-helix bundle fold with up−up−down−down topology and contains intramolecular disulfide bonds. Since an X-ray or NMR structure for OM is not currently available, a homology model for OM was determined from the X-ray structures of human growth hormone (hGH), LIF, and G-CSF where the alignment was based on secondary structure instead of sequence. The OM secondary structure was determined from NMR structural data, and the secondary structures for hGH, LIF, and G-CSF were obtained from the reported X-ray structures. The resulting homology model was refined using sequential NOE distance ^{13}C restraints, chemical shift information, and a conformational database.

The structure of a protein is an essential requirement to obtain a detailed understanding of its biological function and is a necessity for structure-based drug design (*1*, *2*). Unfortunately, current NMR and X-ray techniques incur a large time commitment on the order of 6 months to 1 year to solve the high-resolution structure of a protein (*3*, *4*). Even though there has been a tremendous increase in the number of protein structures solved by NMR or X-ray crystallography, the number of new gene sequences identified far out-pace the rate of protein structure determination (*5*). Recent analysis of the Protein Data Bank indicates that there exists a small number of protein folds which describes reasonably well the 4000 known protein structures (*6*). This suggests that while new protein structures or folds will continue to be discovered, there exists a significant probability that a new protein's fold might be reasonably described by a current structure in the database.

Since the time required to determine a protein homology model is significantly shorter than an NMR or X-ray structure, creating homology models may prove to be the most efficient means of providing structural data for the ever increasing number of new protein sequences (*7*−*11*). The typical approach to homology modeling has relied heavily on sequence similarity between the new protein and target protein(s). The accuracy of the resulting homology model decreases significantly as the sequence homology between the structures drops below 40%. This is clearly problematic since a number of protein families have very divergent sequence homology (20% or less) while maintaining the same overall fold. Attempts have been made to use predictive methods to identify regions of secondary structure to generate alignments for homology modeling, but this approach is lacking since the validity of the resulting structure is dependent on both the combined accuracy of the structure prediction method and the homology modeling protocol. Nevertheless, the concept of using secondary structure alignment instead of sequence alignment to thread a homology model represents a better approach because of the higher information content.

While determining the high-resolution structure of a protein by NMR is still an arduous task, the sequence assignment and resulting identification of secondary structure elements can be accomplished relatively quickly by using $^{15}N/^{13}C$-labeled proteins and standard triple-resonance and 3D-edited NMR techniques. Thus an experimental approach for identifying accurate protein secondary structure elements can be combined with homology modeling to provide a rapid structure of the protein.

The application of NMR-derived structural information to validate the resulting homology model of various proteins has been shown to be quite beneficial. One method used NMR to measure histidine pK_a values to verify that the histidine environment is consistent with the homology model (*12*). Additionally, it has been shown that experimentally determined secondary and tertiary structure NOE-derived distance restraints from a previously solved solution NMR structure may be mapped onto a homologous protein sequence (37% identity) on the basis of sequence alignment and used in a refinement protocol for a homology model. This was demonstrated by combining the experimental restraints of the previously solved *Escherichia coli* acyl carrier protein NMR structure with its spinach homologue to refine the spinach structure (*13*). Also, experimentally derived protein−ligand NOE restraints may provide critical information for refining the homology model's active site as demonstrated by incorporating distance restraints between the human cytochrome P_{450} 2D6 heme iron and the

* Corresponding author.
‡ Wyeth-Ayerst Research.
§ Current address: Albany Molecular Research, Inc., 21 Corporate Circle, Albany, NY 12203.
|| Immunex Corp.
⊥ Current address: Zymo Genetics, Inc., 1201 Eastlake Ave. East, Seattle, WA 98102.

S0006-2960(97)02744-X CCC: $15.00
Published on Web 07/09/1998

codeine ligand on the basis of paramagnetic relaxation times (*14*). Similarly, the proximity of the heme in horseradish peroxidase to neighboring residues deduced by NMR assignments allowed for identification of an appropriate homology model (*15*, *16*). In all cases the homology model of the desired protein was determined by standard sequence alignment techniques where the NMR data were applied to validate the initial model.

Here we describe the first direct application of NMR data to determine a homology model for oncostatin M (OM) (*17*). The secondary structure determined by NMR was used to guide the sequence alignment of OM to hGH, LIF, and G-CSF because the sequence identity was only 20% to the closest protein, LIF. Structure refinement of the model incorporated only sequential NOE distance restraints, ^{13}C chemical shift information, and a conformational database.

OM is a member of the cytokine family that includes interleukin-4 (IL-4), interleukin-6 (IL-6), leukemia inhibitory factor (LIF), and granulocyte-colony stimulating factor (G-CSF). This cytokine family of proteins has been shown to fold as a four-helix bundle with characteristic up−up−down−down helix topology (*18*−*21*) and regulate the proliferation and differentiation of a variety of cell types (*22*, *23*). These shared functions are mediated through the interaction of the signal transducing membrane glycoprotein, gp130, with the extracellular domains of these cytokine receptors (*24*−*26*). Additional biological activities attributed to OM include the inhibition of several tumor cell lines (*18*, *27*), inhibition of embryonic stem cell differentiation (*28*), a regulator of endothelial cells (*29*), and the growth stimulation of several fibroblast cell lines (*18*, *27*). OM has also been shown to be a mitogen for AIDS-related Kaposi's sarcoma cells (*30*, *31*) where it functions as an autocrine growth factor (*32*).

MATERIALS AND METHODS

NMR. The resonance assignments and solution secondary structure for OM determined by NMR were described in detail by Hoffman et al. (*17*). In this previous work, a combination of typical NOE patterns in the ^{15}N NOESY−HMQC spectrum and ^{13}Cα secondary structure chemical shifts was used to determine that the protein is composed of a four-helix bundle (residues 18−45, 78−98, 113−140, and 168−194) with three intervening helical regions (residues 51−57, 66−71, and 103−110). This is consistent with the four-helix bundle motif found for this cytokine family. Regions of α-helices were identified by stretches of sequential NH−NH NOEs and by positive ^{13}Cα secondary chemical shifts (>3 ppm). A total of 240 sequential NOE distance restraints, 173 ϕ and ψ dihedral restraints, 178 Cα secondary shifts, and 141 Cβ secondary shifts where used during the refinement. Because of the ill-behaved nature of the OM NMR sample caused by glycosylation, aggregation, and autolysis, additional structural information for the homology model or for a detailed NMR structure was not available.

Homology Model. The homology model of OM was based on the multiple sequence alignments to bovine OM, LIF, G-CSF, and hGH, which were then checked to determine that the secondary structure from the OM NMR data and the X-ray crystal structures for LIF (1LKI), G-CSF (1RHG), and the hGH (1HUW) aligned properly. The Pileup program (Genetics Computer Group, Madison, WI) was used to produce the primary sequence alignment. The low homology in some regions clearly was problematic, and the pileup method worked well at improving the alignment such that the secondary structure was matched among structures and that the very clear primary sequence similarity and identity matched at the end of the protein chains. The Pileup program produced a multiple sequence alignment such that NMR secondary structural evidence aligns with the known secondary structure of LIF, G-CSF, and hGH. The resulting sequence alignments are shown in Figure 1.

An initial homology model was built using WHATIF (*33*) for those regions where LIF aligned with the OM NMR secondary structure. The PIRPSQ procedure in WHATIF was used to copy the coordinates of conserved amino acids and to mutate the other amino acids. WHATIF uses statistical preferences to provide side-chain conformations and removes close contacts by conformational searching. Initial gaps in the OM structure were filled by copying loops from the G-CSF or hGH crystal structure based on alignment with the OM NMR secondary structure. Again, any necessary side-chain mutations were then made by the WHATIF program. An isolated annealing procedure was used on the two largest gaps (residues 125−145 and 59−78) to overcome problems with initial conformations of these long loops.

The loop annealing procedure was adapted from the work of Fidelis et al. (*34*). The amino acids in loops 125−145 and 59−78 were extracted with the terminal amino acids serving as anchor points. The loops were simulated alone in a molecular dynamics simulation with the relevant torsion and distance restraints for the two regions extracted from the NMR data. The backbone atoms of the first and last two amino acids were fixed during the simulated annealing run, which allowed for the coordinates for each loop to be pasted back into the full structure. This procedure provided a very rapid solution to the structure of the loops and solved the problem of inadequate fragments from the Brookhaven Protein Data Bank.

Annealing Procedure. For each simulated annealing run, only a subset of atoms in the OM homology model were allowed to move while all other atoms were fixed. The weight on the torsion and distance restraints was started at 1 kcal mol^{-1} Å^{-2} and increased to 50 kcal mol^{-1} Å^{-2} over the course of five annealing runs. The simulated annealing protocol at each restraint value was as follows: (1) Minimize the atoms free to move for 500 steps using the conjugate gradient method (restarting every 50 steps); (2) heat the system with molecular dynamics to 100 or 500 K (five 1 ps simulations incrementing the target temperature 20 or 100 K for each step); (3) cool the system with molecular dynamics (five 1 ps simulations decrementing each target temperature 20 or 100 K except for the last cooling step which was targeted at 10 K); and (4) minimize for 1250 steps restarting every 50 steps). The process for a set of free atoms was repeated starting with low restraint energies until the restraint energies were no longer improving or until any structural artifacts were eliminated (such as incorrect proline puckering). Temperature control on the molecular dynamics simulations was accomplished by coupling to an external temperature bath. Final temperatures were set to 500 K when annealing small sections of the protein and

```
          1                                                  50
Bovine OM MGAQRMQRTL LSLVLRLLLL CTVATGKCSG K..YHE.LLL QLQRQ.ADLM
Human OM  ......... ......... ..AAIGSCSK E..YRV.LLG QLQKQ.TDLM
LIF       ......... ....SPLPIT PVNATCAIRH P..CHGNLMN QIKNQLAQLN
G-CSF     ......... .........T PLGPASSLPQ S..FLLKCLE QVRKIQGDGA
hGH       ......... .........F PTIPLSRLAD NAWLRADRLN QL......AF

          51                                                 100
Bovine OM QDPSTLLDPY IHLQGLHSP. VLQEHCRERP GDFPSEDALW RLSRQDFLQT
Human OM  QDTSRLLDPY IRIQGLDVP. KLREHCRERP GAFPSEETLR GLGRRGFLQT
LIF       GSANALFISY YTAQGEPFPN NVEKLCAPNM TDFPSFHG.. .......NGT
G-CSF     ALQEKLCATY KLCHPEEL.. VLLGHSLGIP WAPLSSCPSQ ALQLAGCLSQ
hGH       DTYQEFEEAY IPKEQIHSFW WNPQTSLCPS ESIPTPSNKE ETQQKSNLEL

          101                                                150
Bovine OM LNTTLGLILR MLSALQQDLP EAAH...... .......... QQAEMNVRGF
Human OM  LAATLGAVLH RLADLEQRLP KAQDLER..S GLNIEDLEKL QMARPNILGL
LIF       EKTKLVELYR MVAYLSASLT NITRDQKVLN PTAVSLQVKL NATIDVMRGL
G-CSF     LHSGLFLYQG LLQALEGISP ELGP...... ........TL DTLQLDVADF
hGH       LRISLLLIQS WLEPV.QFLR SVFANSLVYG ASDSNVYDLL KDLEEGIQTL

          151                                                200
Bovine OM GNNIHCMAQL LRGSSDPKAA EPTQPGPGPT .PLPPTPPSS TFQRKLRNCG
Human OM  RNNIYCMAQL LDNS...DTA EPTKAGRGAS QPPTPTPASD AFQRKLEGCR
LIF       LSNVLC.... .......RLC NKYRVGHVDV PPVPDHSDKE AFQRKKLGCQ
G-CSF     ATTIWQQMEE LG......MA PALQPTQGAM PAFA.....S AFQRRAGGVL
hGH       MGRLEALLKN YGLLYCFNKD MSKVSTYLRT VQCRSVEGSC GF........

          201                                                250
Bovine OM FLRGYHRFMR TAGQVLRGWG ERQGRSRRHS PCRALKRGAR RTQPFPEIRR
Human OM  FLHGYHRFMH SVGRVFSKWG ESPARSR... .......... ..........
LIF       LLGTYKQVIS VVVQAF.... .......... .......... ..........
G-CSF     VASHLQSFLE VSYRVLRHLA QP........ .......... ..........
hGH       .......... .......... .......... .......... ..........

          251           267
Bovine OM LAPRGQPPGS LWGAPAR
Human OM  ......... .......
LIF       ......... .......
G-CSF     ......... .......
hGH       ......... .......
```

FIGURE 1: Sequence alignment of bovine OM, human OM, LIF, G-CSF, and hGH based on the NMR-derived secondary structure for OM and the X-ray secondary structure for LIF, G-CSF, and hGH. Helical regions are shown in bold italics.

the 100 K value when large segments of the protein were allowed to move. The AMBER4.0 force field (*35*, *36*) was used in the annealing calculations using the IMPACT program (*37*).

Refinement. The procedure used to refine the structure and achieve agreement with experimental measurements was developed in a highly empirical method since standard methods of annealing at high temperature did not work because of the lack of experimental restraints on the tertiary structure.

Two sets of torsion restraints were used during the refinement of the OM structure. Initially torsion and distance restraints were based on defined α-helix (ϕ and ψ restrained to 60 ± 50°) and β-strand regions (ϕ restrained to 135 ± 50° and ψ restrained to be −140 ± 50°) with additional distance restraints between i,i+4 O and N atoms and sequential restraints from the NMR experiments. This resulted in 240 NMR restraints and 76 heavy atom restraints. In addition, two intramolecular disulfide bonds were incorporated between C6−C127 and C49−C167 as previously determined for OM (*38*). The disulfide bonds were modeled initially by distance restraints between the sulfurs of the two cysteines as well as Cα−Cα distances and other intra- and interresidue heavy atom distances so that the disulfide bonds could be formed gently while maintaining the correct stereochemistry of the cysteines. The deletions and insertions in the structure were handled by placing additional Cα−Cα sequential restraints (3.7−3.9 Å) and sequential Ω restraints (CO_i to HN_{i+1}; 3.0−3.2 Å). These additional distance restraints were found to help during the annealing process by properly closing gaps and maintaining chirality and transpeptide bonds.

The initial set of torsion restraints based on standard secondary structure geometry was found to be inadequate to drive the protein structure toward a solution that could be used in an X-PLOR run with $^{13}C\alpha/^{13}C\beta$ chemical shift restraints. Instead, two sets of torsion restraints were found to be helpful during refinement: an experimentally derived set and a homology-based set. To derive the experimental set the values of torsion restraints and their ranges were expanded to include the ϕ/ψ torsions predicted by secondary $^{13}C\alpha/^{13}C\beta$ chemical shifts (*39*). Most of the secondary $^{13}C\alpha/^{13}C\beta$ chemical shifts agreed with the secondary structure assignments. In cases where there was disagreement (or no previous assignment) the center of the torsion restraints and the allowed range were adjusted so that the final restraints spanned the ϕ/ψ space predicted by the $^{13}C\alpha/^{13}C\beta$ chemical shifts. Because this set of restraints was distant from the homology model starting structure in several amino acids, another set of restraints was helpful in the early stages of refinement. The homology-based set of restraints was derived on the basis of the LIF structure (1LKI) (*21*). The ϕ/ψ values from LIF were calculated and then compared to the experimental restraints for the homologous residues in the model. Torsion restraints were generated to span the homologous LIF values and the ϕ/ψ space predicted by the $^{13}C\alpha/^{13}C\beta$ chemical shifts. This set of restraints was only used in the first pass of annealing to move the initial model away from the homology model toward the final experimental numbers. After the initial refinements only the NMR-determined restraints were used.

As stated previously, because of the lack of tertiary restraints high-temperature simulated annealing to refine the homology model and generate a structure consistent with the experimental restraints was problematic. To overcome this difficulty, the initial homology-modeled structure was refined using the two sets of restraints using many different runs where only selected atoms were free to move. In essence, this provided a stepwise approach to overcome local conformational problems while maintaining a proper overall fold to the protein. The stepwise annealing procedure is described below.

The first run was started without disulfide bonds (but with distance restraints that directed the formation of the disulfide bonds) and with the LIF homology model set of torsion restraints. In this round the α-helical bundle backbone atoms remained fixed and only loops were annealed. After no further progress was obtained in satisfying restraint energies, the next run was started with the disulfide bonds formed and using the final set of experimental torsion restraints. Initially, only the Cα atoms of the helix bundle were fixed. The refinement was continued until no further improvement in restraint violations was observed.

The next run of the simulation addressed problematic sections containing prolines, regions where there had been gaps in the alignment and regions where unusual $^{13}C\alpha/^{13}C\beta$ chemical shifts were observed. Only atoms in each residue of these regions were allowed to move during the simulation. It was found that it was necessary to run the simulation at temperatures up to 500 K to anneal these sections. After no further progress was obtained in satisfying restraint violations, residues 1−24, 30−37, 90−100, and 170−190 were allowed to move. Runs were repeated until the agreement with experimental restraints was not improving and until artifactual changes in the protein structure were no longer being repaired. This procedure allowed the proper solution of restraints in these regions and to improve the agreement of the structure with known preferences of proteins as determined by WHATCHECK. It was found that if the entire helix bundle was permitted to move as the restraints in the N-terminal region were being annealed, then unrealistic refolding of the bundle was observed [as determined by checks with PROCHECK, WHATCHECK, and visual inspection (*33*, *40*)]. A group of amino acids were identified throughout the structure that had remaining structural difficulties, and these were resolved by minimization of these amino acids as a group without any high-temperature molecular dynamics. After the final structure was checked with the Eisenberg 3D−1D method, manual intervention was used to adjust the first and fourth helix in order to twist it as a unit (*41*, *42*). This structure was then annealed to 100 K in order to fix any artifacts introduced. The final structure was then minimized using the program X-PLOR (*43*), adapted to incorporate secondary $^{13}C\alpha/^{13}C\beta$ chemical shift restraints (*44*) and a conformational database potential (*45*, *46*).

RESULTS AND DISCUSSION

Description of the Structure. A ribbon diagram for the refined NMR-based homology model of OM is shown in Figure 2. The main feature of the protein's structure is the four long helices forming an up−up−down−down four helix bundle connected by intervening short or very long loop sections typical of this class of cytokines. In addition to the four-helix bundle there are three intervening short helical regions, one corresponding to a short turn region between helices B and C and two occurring in a long loop region between helices A and B. OM contains two disulfide bonds which play a significant role in the packing of the protein. The disulfide bond between Cys 6 and Cys 127 assists in packing helices A and C while the disulfide bond between Cys 49 and Cys 167 connects helix D with the first short helix in the long loop region between helices A and B.

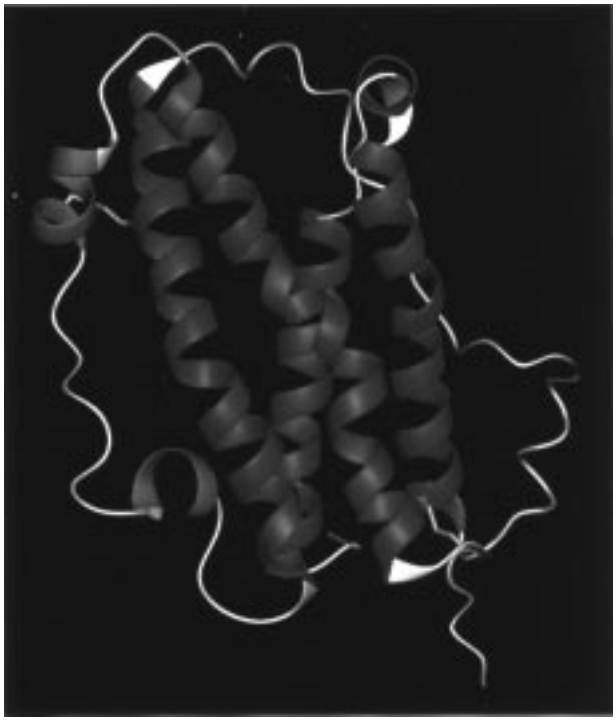

FIGURE 2: Ribbon drawing of the NMR-based homology model of OM where the four-helix bundle is colored purple, the intervening helical regions are colored red, and the disulfide bonds are colored yellow.

It is important to note that regions of helices A and D are partially distorted in the homology model of OM. The addition of $^{13}C\alpha/\beta$-based torsion restraints in the refinement protocol produced a structure compatible with OM's secondary structure $^{13}C\alpha$ and $^{13}C\beta$ chemical shift assignments during final minimization. The positive ψ regions in these helices were not easily achieved with the level of structural information available. It appears that solving for the positive ψ values allows for many unlikely structures (e.g., highly exposed cores of the protein). But the fact that the sequential assignments also show a breakdown of the helical regions suggests that the positive ψ values are correctly assigned although the alignment to LIF may be incorrect in these areas.

Structure Validation. Initial confidence in the 3D model was obtained because of the relative ease of forming disulfide bonds and the agreement of the alignment with the observed NMR results for OM and other structures in its family. Significant uncertainty in the structure obtained in this work occurs in areas similar to those of crystal structures. In all the structures of the four-helix bundle family there is significant disorder in the long loop regions where some atomic positions are not reported. Regions of extreme disorder in the X-ray structures where no atomic positions are reported correspond to residues 130−154 in hGH (the long loop region between helices C and D) and to residues 65−71 and 127−136 in G-CSF (corresponding to the long loop regions between helices A and B and helices C and D, respectively). Similarly, significant NMR restraints are not obtained for large sections of the OM loop structure, and therefore these atomic positions are not considered accurate, whereas the positions of the helical regions are likely to be similar to those obtained if a full NMR or crystal structure could be obtained. This is evident by the plot of the error function obtained from the ERRAT program per residue

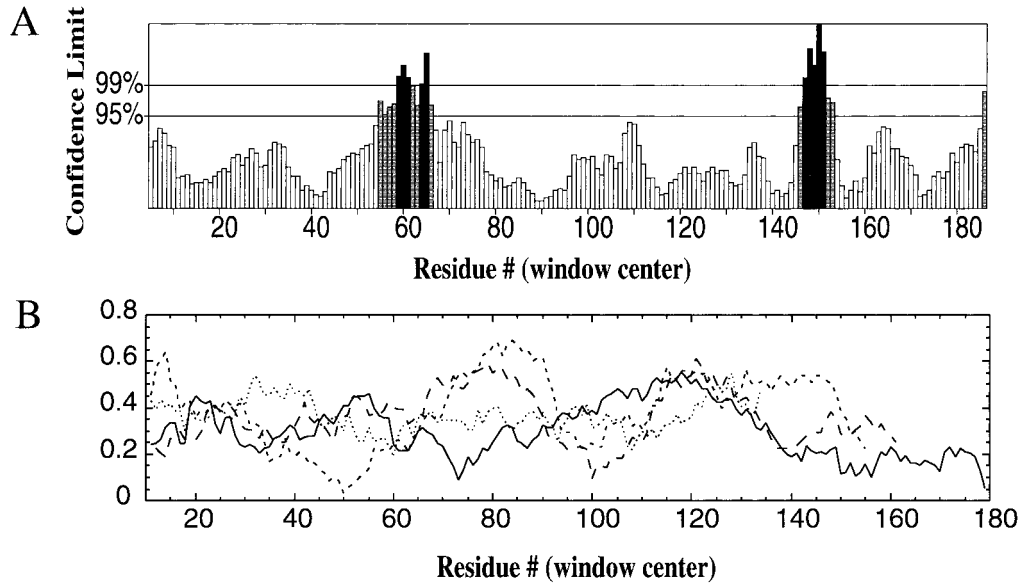

FIGURE 3: (A) Plot of the error function per residue for the NMR-based homology model of OM determined by the ERRAT (*47*) program for nonbonded atom−atom interactions. (B) Plot of the 3D profile per residue for the NMR-based homology model of OM (—), 1HUW (---), 1LKL (− −), and 1RHG (•••). The vertical axis gives the average 3D−1D score for residues in a nine-residue sliding window, the center of which is at the sequence position indicated by the horizontal axis (*41*, *42*).

(Figure 3), where large errors in the nonbonded atom−atom interactions for the long loop regions (residues 56−78 and 141−153) are observed (*47*). The packing of hydrophilic amino acids in the core of the bundle is consistent with other four-helix bundles. It is also found that the electrostatic picture of the protein is similar to others in the family.

To further substantiate the validity of the OM homology model, the structure was analyzed using PROCHECK and WHATIF (*33*, *40*). For the PROCHECK statistics, an overall *G*-factor of −0.21, hydrogen bond energy of 1.1, and only 0.27 bad contacts per 100 residues are observed, which is consistent with a good quality structure comparable to a high-resolution X-ray structure. Similarly, WHATIF analysis indicates a packing score of −1.69 with side-chain rotamer and backbone scores of 0.64 and −0.37, respectively, consistent with a properly folded protein structure. Additional evidence for the reliability of the OM homology model is seen in the Ramachandran ϕ, ψ plot in Figure 4. Most of the backbone torsion angles for non-glycine residues lie within expected regions of the Ramachandran plot; 86.3% of the residues lie within the most favored region of the Ramachandran ϕ, ψ plot, 13.0% in the additionally allowed regions, and 0.6% in the generously allowed region. Further validity of the OM homology model is obtained from a 3D profile of the structure (*41*, *42*). The calculated *Z*-value for the OM model is 51.7, which is consistent with the *Z*-values for LIF (53.4), G-CSF (50.0), and hGH (59.3). The apparently low *Z*-values for these proteins are probably attributed to the ill-defined long loop regions. A comparison of the profile window plot between the OM model and the reference proteins is shown in Figure 3. Again, a reasonable similarity is obtained for the OM model values compared to the reference proteins. The quality of the OM structure is also evident by the very small deviations from idealized covalent geometry.

Comparison to Previous Homology Model. It is important to note that the homology model reported here for OM differs significantly from a model previously reported, based strictly

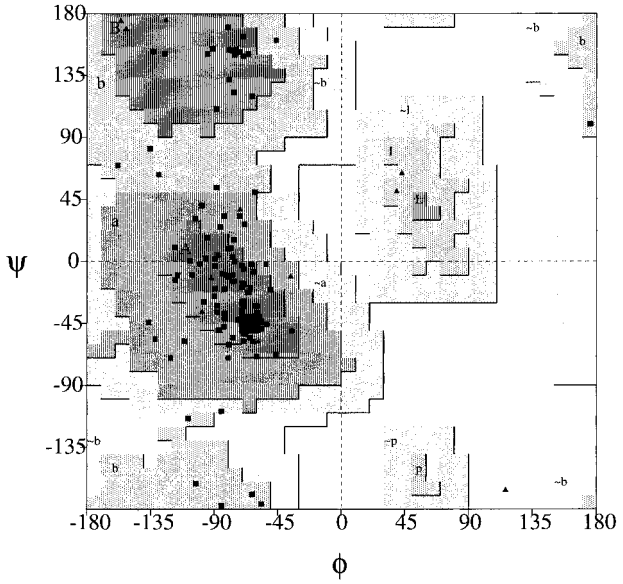

FIGURE 4: Ramachandran ϕ, ψ plot for the NMR-based homology model of OM. The glycine residues are plotted as solid triangles.

on a sequence alignment with porcine growth hormone (*19*, *48*, *49*). In particular, an eight-residue shift occurs for helix B, and the short intervening helical regions are not observed in the original homology model. This difference arises primarily from the choice of the porcine growth hormone sequence as the target protein for the homology modeling. By relying on the NMR secondary structure information to base a sequence alignment with a number of proteins from the OM family (hGH, LIF, and G-CSF), a more reliable OM homology model is obtained. The resulting structural differences between the two homology models exemplify the inherent difficulties in deriving a reliable model based strictly on a sequence alignment below 40% sequence identity while demonstrating the added benefit of incorporating experimental NMR data to assist in homology modeling.

This is particularly evident for the OM family of homologous proteins. While OM and other members of this

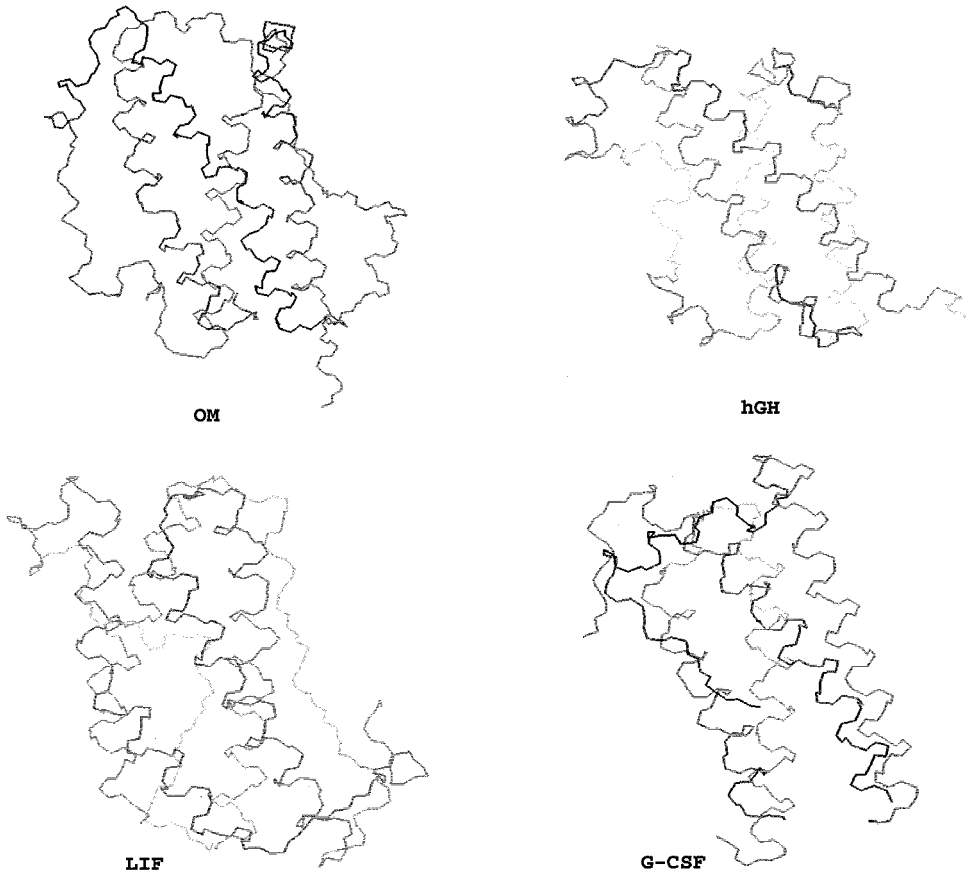

FIGURE 5: Structural comparison of the NMR-based homology model of OM with the X-ray structures of hGH, LIF, and G-CSF.

structural family adopt the general topology of a four-helix bundle with characteristic up−up−down−down helix topology, the details of each protein structure vary distinctly. Particularly, the relative lengths of the helices and intervening loops, the number of disulfide bonds, and more strikingly the presence of β-sheets or helical regions in the intervening loops further delineate the members of this structural family. Given the relatively low sequence homology in this structural class, it appears to be very difficult to a priori determine which protein structure to use as a basis for a homology model from sequence alignment only. This is clearly evident by the incorrect OM homology model using the porcine growth hormone. In fact, in our experience with OM, it appears that the hybrid approach of using multiple sequence alignments of proteins with similar secondary structure composition to align with the NMR secondary structure to be the best solution.

Biological Relevance. The ultimate test for the accuracy of a homology model is consistency with biological activity. OM is a member of a structural family that includes human growth hormone (hGH), leukemia inhibitory factor (LIF), and granulocyte-colony stimulating factor (G-CSF). Major differences between these proteins include the relative length of the helices in the four-helix bundle, the number of disulfide bonds, the number of intervening helices, and the overall size of the protein. A comparison of the folds for these three proteins with OM is shown in Figure 5. Structural and biological evidence indicates that hGH stimulates a response by inducing receptor dimerization, where hGH contains two distinct receptor binding sites (*50*−*53*). It has been postulated that other growth factors and cytokines, including OM, also induce a signal cascade by receptor dimerization. Site-directed mutagenesis of OM has indicated a number of residues which are critical or essential for OM activity (184−190, 22−3, 44−47, and the Cys 49−Cys 167 disulfide bond) (*38*, *54*). A molecular surface for the OM homology model has been calculated with GRASP (*55*) where critical or essential residues for activity are colored (Figure 6). This map indicates the presence of two distinct surfaces on OM from the mutagenenous studies and is consistent with OM binding two receptors for activity. These results provide additional support for the validity of the OM homology model.

CONCLUSION

An NMR-based homology model for OM is presented where the secondary structure determined by NMR was used to guide the sequence alignment of OM to hGH, LIF, and G-CSF and the structure was refined using sequential distance restraints, ^{13}C chemical shift information, and a conformational database. The use of the latest techniques in X-PLOR increases our confidence that the model is substantially correct in the secondary structure regions and that secondary structural elements are likely to be accurately positioned with respect to each other as evident by the structural validation by PROCHECK, WHATIF, and ER-RAT. Consistency between the homology model and the mutagenesis data with the proposed receptor dimerization further supports the homology model.

The model was refined using recently reported methods based on ^{13}C chemical shifts and protein quality checks. It

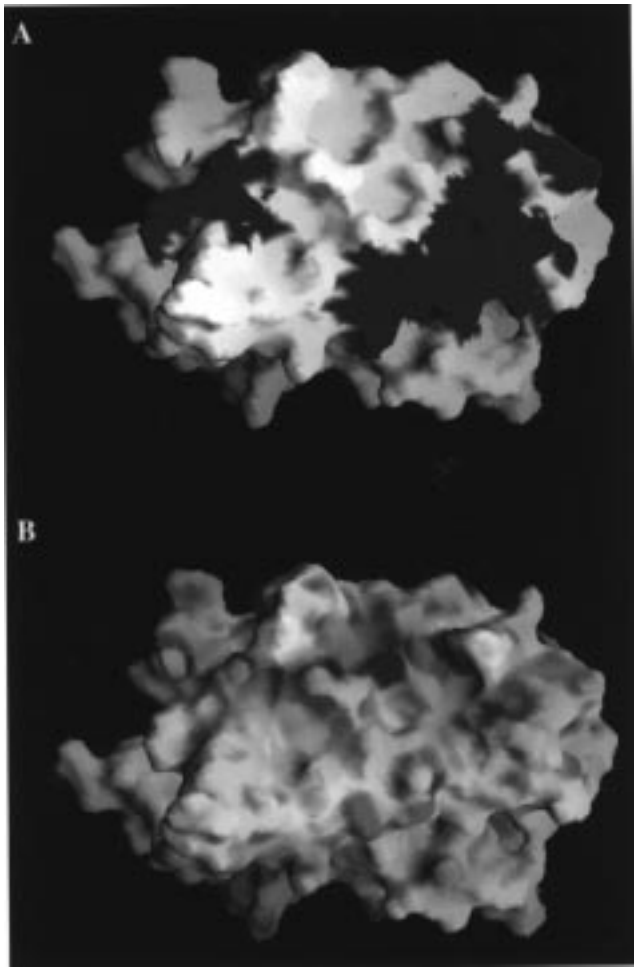

FIGURE 6: GRASP surface determined for the NMR-based homology model of OM. (A) Surface colored blue by essential residues from mutagenesis studies. (B) Surface colored by electrostatic potential where blue is positive and red negative.

was found that converting ^{13}C chemical shifts to torsional restraints allowed the model to shift in unusual regions toward the ^{13}C values and away from the homology-based backbone structure derived from hGH, LIF, and G-CSF. We believe that the structure presented is significantly more accurate than homology modeling by itself and could not have been obtained simply by using sequence alignment.

Homology models derived from coupling simple NMR data with low homology structures should be useful in determining global folds of proteins, aiding in phase determination or model building as new experimental data is obtained, and design of mutagenesis experiments and protein engineering.

REFERENCES

1. Whittle, P. J., and Blundell, T. L. (1994) *Annu. Rev. Biophys. Biomol. Struct. 23*, 349−375.
2. Blundell, T. L. (1996) *Nature 384*, 23−26.
3. Clore, G. M., and Gronenborn, A. M. (1991) *Science 252*, 1390−1399.
4. Clore, G. M., and Gronenborn, A. M. (1994) *Protein Sci. 3*, 372−390.
5. Bowie, J. U., Luthy, R., and Eisenberg, D. (1991) *Science 235*, 164−170.
6. Holm, L., and Sander, C. (1996) *Science 273*, 595−607.
7. Greer, J. (1990) *Proteins: Struct., Funct., Genet. 7*, 317−334.
8. Siezen, R. J., de Vos, W. M., Leunissen, J. A. M., and Dijkstra, B. W. (1991) *Protein Eng. 4*, 719−737.
9. Hilbert, M., Bohm, G., and Jaenicke, R. (1993) *Proteins: Struct., Funct., Genet. 17*, 138−151.
10. Bajorath, J., Stenkamp, R., and Aruffo, A. (1993) *Protein Sci. 2*, 1798−1810.
11. Benner, S. A., Cannarozzi, G., Gerloff, D., Turcotte, M., and Chelvanayagam, G. (1997) *Chem. Rev. 97*, 2725−2843.
12. Smith, D. K., Treutlein, H. R., Maurer, T., Owczarek, C. M., Layton, M. J., Nicola, N. A., and Norton, R. S. (1994) *FEBS Lett. 350*, 275−280.
13. Oswood, M. C., Kim, Y., Ohlrogge, J. B., and Prestegard, J. H. (1997) *Proteins: Struct., Funct., Genet. 27*, 131−143.
14. Modi, S., Paine, M. J., Sutcliffe, M. J., Lian, L. Y., Primrose, W. U., Wolf, C. R., and Roberts, G. C. K. (1996) *Biochemistry 35*, 4540−4550.
15. Loew, G. H., Du, P., and Smith, A. T. (1995) *Biochem. Soc. Trans. 23*, 250−256.
16. Smith, A. T., Du, P., and Loew, G. H. (1995) *NATO ASI Ser., Ser. C 457*, 75−93.
17. Hoffman, R. C., Moy, F. J., Price, V., Richardson, J., Kaubisch, D., Frieden, E. A., Krakover, J. D., Castner, B. J., King, J., March, C. J., and Powers, R. (1996) *J. Biomol. NMR 7*, 273−282.
18. Zarling, J. M., Shoyab, M., Marquardt, H., Hanson, M. B., Lioubin, M. N., and Todaro, G. J. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 9739−9743.
19. Rose, T. M., and Bruce, A. G. (1991) *Proc. Natl. Acad. Sci. U.S.A. 88*, 8641−8645.
20. Powers, R., Garrett, D. S., March, C. J., Frieden, E. A., Gronenborn, A. M., and Clore, G. M. (1993) *Biochemistry 32*, 6744−6762.
21. Robinson, R. C., Grey, L. M., Staunton, D., Vankelecom, H., Vernallis, A. B., Moreau, J. F., Stuart, D. I., Heath, J. K., and Jones, E. Y. (1994) *Cell 77*, 1101−1116.
22. Tanigawa, T., Nicola, N., McArthur, G. A., Strasser, A., and Begley, C. G. (1995) *Blood 85*, 379−390.
23. Bruce, A. G., Hoggatt, I. H., and Rose, T. M. (1992) *J. Immunol. 149*, 1271−1275.
24. Sporeno, E., Paonessa, G., Salvati, A. L., Graziani, R., Delmastro, P., Ciliberto, G., and Toniatti, C. (1994) *J. Biol. Chem 269*, 10991−10995.
25. Liu, J., Modrell, B., Aruffo, A., Scharnowske, S., and Shoyab, M. (1994) *Cytokine 6*, 272−278.
26. Boulton, T. G., Stahl, N., and Yancopoulos, G. D. (1994) *J. Biol. Chem. 269*, 11648−11655.
27. Horn, D., Fitzpatrick, W. C., Gompper, P. T., Ochs, V., Bolton-Hanson, M., Zarling, J., Malik, N., Todaro, G. J., and Linsley, P. S. (1990) *Growth Factors 22*, 157−165.
28. Rose, T. M., Weiford, D. M., Gunderson, N. L., and Bruce, A. G. (1994) *Cytokine 6*, 48−54.
29. Brown, T. J., Rowe, J. M., Shoyab, M., and Gladstone, P. (1990) *UCLA Symp. Mol. Cell. Biol., New Ser. 131*, 195−206.
30. Miles, S. A., Martinez-Maza, O., Rezai, A., Magpantay, L., Kishimoto, T., Nakamura, S., Radka, S. F., and Linsley, P. S. (1992) *Science 255*, 1432−1434.
31. Nair, B. C., DeVico, A. L., Nakamura, S., Copeland, T. D., Chen, Y., Patel, A., O'Neil, T., Oroszlan, S., Gallo, R. C., and Sarngadharan, M. G. (1992) *Science 255*, 1430−1432.
32. Cai, J., Gill, P. S., Masood, R., Chandrasoma, P., Jung, B., Law, R. E., and Radka, S. F. (1994) *Am. J. Pathol. 145*, 74−79.
33. Vriend, G., and Sander, C. (1993) *J. Appl. Crystallogr. 26*, 47−60.
34. Fidelis, K., Stern, P. S., Bacon, D., and Moult, J. (1994) *Protein Eng. 7*, 953−960.
35. Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, I., T. E., Ferguson, D. M., Seibel, G. L., Singh, U. C., Weiner, P. K., and Kollman, P. A. (1995) *AMBER*, version 4.1, University of California, San Francisco.

36. Cornell, W. D., Cieplak, P., Baylyl, C. I., Gould, I. R., Merez, J., K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995) *J. Am. Chem. Soc. 117*, 5179.
37. Kitchen, D. B., Hirata, F., Westbrook, J. D., Levy, R. M., and Yarmush, M. (1990) *J. Comput. Chem. 11*, 1169−1180.
38. Kallestad, J. C., Shoyab, M., and Linsley, P. S. (1991) *J. Biol. Chem. 266*, 8940−8945.
39. Spera, S., and Bax, A. (1991) *J. Am. Chem. Soc 113*, 5490−5492.
40. Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993) *J. Appl. Crystallogr. 26*, 283−291.
41. Bowie, J. U., Luthy, R., and Eisenberg, D. (1991) *Science 253*, 164−170.
42. Luthy, R., Bowie, J. U., and Eisenberg, D. (1992) *Nature 356*, 83−85.
43. Brunger, A. T. (1993) *X-PLOR Version 3.1 Manual*, Yale University, New Haven, CT.
44. Kuszewski, J., Qin, J., Gronenborn, A. M., and Clore, G. M. (1995) *J. Magn. Reson., Ser. B 106*, 92−96.
45. Kuszewski, J., Gronenborn, A. M., and Clore, G. M. (1996) *Protein Sci. 5*, 1067−1080.
46. Kuszewski, J., Gronenborn, A. M., and Clore, G. M. (1997) *J. Magn. Reson. 125*, 171−177.
47. Colovos, C., and Yeates, T. O. (1993) *Protein Sci. 2*, 1511−1519.
48. Abdel-Meguid, S. S., Shieh, H. S., Smith, W. W., Dayringer, H. E., Violand, B. N., and Bentle, L. A. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 6434−6437.
49. Bazan, J. F. (1990) *Immunol. Today 11*, 350−354.
50. De Vos, A. M., Ultsch, M., and Kossiakoff, A. A. (1992) *Science 255*, 306−312.
51. Fuh, G., Cunningham, B. C., and Fukunaga, R. (1992) *Science 256*, 1677−1680.
52. Cunningham, B. C., Ultsch, M., and De Vos, A. M. (1991) *Science 254*, 821−825.
53. Boulay, J. L., and Paul, W. E. (1992) *J. Biol. Chem 267*, 20525−20528.
54. Radka, S. F., Kallestad, J. C., Linsley, P. S., and Shoyab, M. (1994) *Cytokine 6*, 55−60.
55. Nicholls, A., Sharp, K., and Honig, B. (1991) *Proteins: Struct., Funct., Genet. 11*, 281.

BI9727444