Protein Science (1993), *2*, 277–289. Cambridge University Press. Printed in the USA.

An automated method for modeling proteins on known templates using distance geometry

SUBHASHINI SRINIVASAN, CARL J. MARCH, AND SUCHA SUDARSANAM
Department of Protein Chemistry, Immunex Corporation, Seattle, Washington 98101

(RECEIVED July 2, 1992; REVISED MANUSCRIPT RECEIVED October 7, 1992)

Abstract

We present an automated method incorporated into a software package, FOLDER, to fold a protein sequence on a given three-dimensional (3D) template. Starting with the sequence alignment of a family of homologous proteins, tertiary structures are modeled using the known 3D structure of one member of the family as a template. Homologous interatomic distances from the template are used as constraints. For nonhomologous regions in the model protein, the lower and the upper bounds for the interatomic distances are imposed by steric constraints and the globular dimensions of the template, respectively. Distance geometry is used to embed an ensemble of structures consistent with these distance bounds. Structures are selected from this ensemble based on minimal distance error criteria, after a penalty function optimization step. These structures are then refined using energy optimization methods.

The method is tested by simulating the α-chain of horse hemoglobin using the α-chain of human hemoglobin as the template and by comparing the generated models with the crystal structure of the α-chain of horse hemoglobin. We also test the packing efficiency of this method by reconstructing the atomic positions of the interior side chains beyond Cβ atoms of a protein domain from a known 3D structure. In both test cases, models retain the template constraints and any additionally imposed constraints while the packing of the interior residues is optimized with no short contacts or bond deformations. To demonstrate the use of this method in simulating structures of proteins with nonhomologous disulfides, we construct a model of murine interleukin (IL)-4 using the NMR structure of human IL-4 as the template. The resulting geometry of the nonhomologous disulfide in the model structure for murine IL-4 is consistent with standard disulfide geometry.

Keywords: distance geometry; disulfide crosslinks; helical cytokines; hematopoietin receptor gene superfamily; homology modeling; interleukin-4; protein folding; template modeling

The three-dimensional (3D) structural similarity between proteins of unknown structures has generally been inferred from their relative distance to one another in a phylogenetic tree constructed from primary sequence information. Homologous proteins with strong sequence similarity are known to have similar tertiary structures. However, proteins with weak sequence similarity, in some cases, show strong structural homology. For example, the crystal structures of CD4 (Ryu et al., 1990; Wang et al., 1990) and PapD (Holmgren & Branden, 1989) show strong structural homology but weak sequence similarity. This suggests that the diversity in tertiary topology is much more limited than the diversity found in the amino acid sequences of proteins. Although much progress has been made in understanding the relationship between primary sequence information and the tertiary structure it encodes, the rules of protein folding remain poorly understood. Thus, modeling of protein structures based on sequence homology remains the primary means of predicting the structure of a protein prior to obtaining actual X-ray crystallography or NMR structural data.

Homology modeling/comparative modeling was used by Browne et al. (1969) to model bovine α-lactalbumin from hen egg-white lysozyme. McLachlan and Shatton (1971) pointed out the structural similarities between α-lytic protease of *Myxobacter* 495 and elastase, which belong to the same family. More recently, a systematic approach to homology/comparative modeling was proposed by Greer (1981) to model members in a protein family when the 3D structures of some members of the family are known. In this method, the known 3D structures of the proteins in a given family are superimposed

Reprint requests to: Subhashini Srinivasan, Department of Protein Chemistry, Immunex Corporation, 51 University Street, Seattle, Washington 98101.

to define the structurally conserved regions in that family. Among the members of a given family, there is considerable variation in the conformations of regions located between two consecutive structurally conserved regions, and thus, these regions are called the variable regions. These variable regions essentially contribute to the identity of a protein in its family. Although the Cartesian coordinates of the template protein are generally successful in constructing the structurally conserved regions for a structurally undefined member of that family, the modeling of the variable regions has been a major problem in this approach. Many suggestions to solve this problem exist in the literature (e.g., Fine et al., 1986; Moult & James, 1986; Snow & Amzel, 1986; Bruccoleri & Karplus, 1987; Chothia et al., 1989; Claessens et al., 1989; Martin et al., 1989). However, all of these methods have their shortcomings, some of which are highlighted below.

The homology method proposed by Greer (1981) along with the knowledge-based approach (Claessens et al., 1989) for modeling variable regions often lead to models with high-energy short contacts between nonbonded atoms. These high-energy contacts are usually between intervariable regions that are grafted from different known protein structures. In cases where the sequence identity in the structurally conserved regions between the template and the model protein is weak, the interior residues are also susceptible to short contacts. Generally, these short contacts are removed by performing rotation around single bonds using interactive graphics, which is a tedious and, at times, impractical procedure. Energy minimization is used to relax strains in a model. However, the minimization procedure leads to structures that are trapped in local minima and relies entirely on the integrity of the starting structure. Knowledge-based methods used to model insertion/deletion sites require a database of similar local folds and are not useful to model proteins on templates that are of unique tertiary topology.

Ponder and Richards (1987) used a combinatorial approach to replace the buried side chains of template to generate a list of sequences that could be packed in a given template. In this method, the backbone of the template protein is held fixed while changing the side chains and their conformers. However, it is well documented that the backbone atoms of proteins folding in the same template are not necessarily superposable (Chothia & Lesk, 1982). In structurally homologous proteins sharing the immunoglobulin (Ig) fold, a twist of nearly 20–30° between the sheets is tolerated, as well as a root mean square (rms) deviation of more than 2–3 Å in the backbone. Additionally, swelling of the entire Ig fold is observed in the structures of PapD (Holmgren & Branden, 1989) and CD4 (Ryu et al., 1990; Wang et al., 1990) when compared to the Ig structures. Thus, homology modeling methods need to accommodate repositioning of backbone atoms.

The tertiary topology of a protein is an invariant with respect to distance space, dictated only by the interaction between atoms in a manner proportional to the distance between all of the atoms in a protein. Thus, it is realistic to compare the homology between protein structures in distance space. Distance geometry methods have been used earlier by Srinivasan et al. (1986) to build a model of bungarotoxin consistent with the topological constraints from two other known homologous structures, along with the other known experimental constraints for bungarotoxin. Havel and Snow (1991) used a similar approach to simulate the structures of several Kazal-type trypsin inhibitors based on sequences with a relatively high percentage of identity and on structures with absolute conservation of disulfide bridge topology. In both cases, it has been shown that the structure conservation constraints are well retained by the model structures.

Here, we describe a systematic approach to homology modeling using distance geometry. We present the algorithm and a set of computer programs that automates the simulation of protein models with strong or weak sequence similarity to a known tertiary template using distance geometry. This method also provides for modeling proteins with large insertion/deletion regions and that contain disulfides that are nonhomologous to the template protein. The various steps in the simulation of murine interleukin (IL)-4 model on the solution structure of human IL-4 are used to illustrate the strengths of our method.

Results

To assess the ability of our method to model homologous proteins with high sequence identity we utilized the α-chain of human hemoglobin (Brookhaven Protein Data Bank [PDB] code: 2HHB) as a template to model the α-chain of horse hemoglobin (PDB code: 2DHB). The horse hemoglobin α-chain bears 84% sequence identity with the human hemoglobin α-chain. From the sequence alignment file, a hybrid coordinate file for the horse hemoglobin subunit was constructed as described in the Methods. A distance constraint file and a hybrid structure file for the model protein was generated by the automated package (described in the Methods) within a few seconds on a Silicon Graphics workstation model 4D/220. The hybrid structure file and the distance constraint file were used to embed 20 structures using DGEOM, a distance geometry program (Blaney et al., 1989) and three structures with the least distance errors were selected. These three structures are shown superimposed on the crystal structures of human and horse hemoglobin (Fig. 1A–C). The rms deviations of the three model structures from the template structure are given in Table 1. Table 1 also includes the rms deviations of horse hemoglobin model structures with that of the crystal structure. The rms deviations between the model structures simulated by our method and the template are observed to be 0.001 Å for both Cα atoms and backbone atoms (Table 1), showing that the model

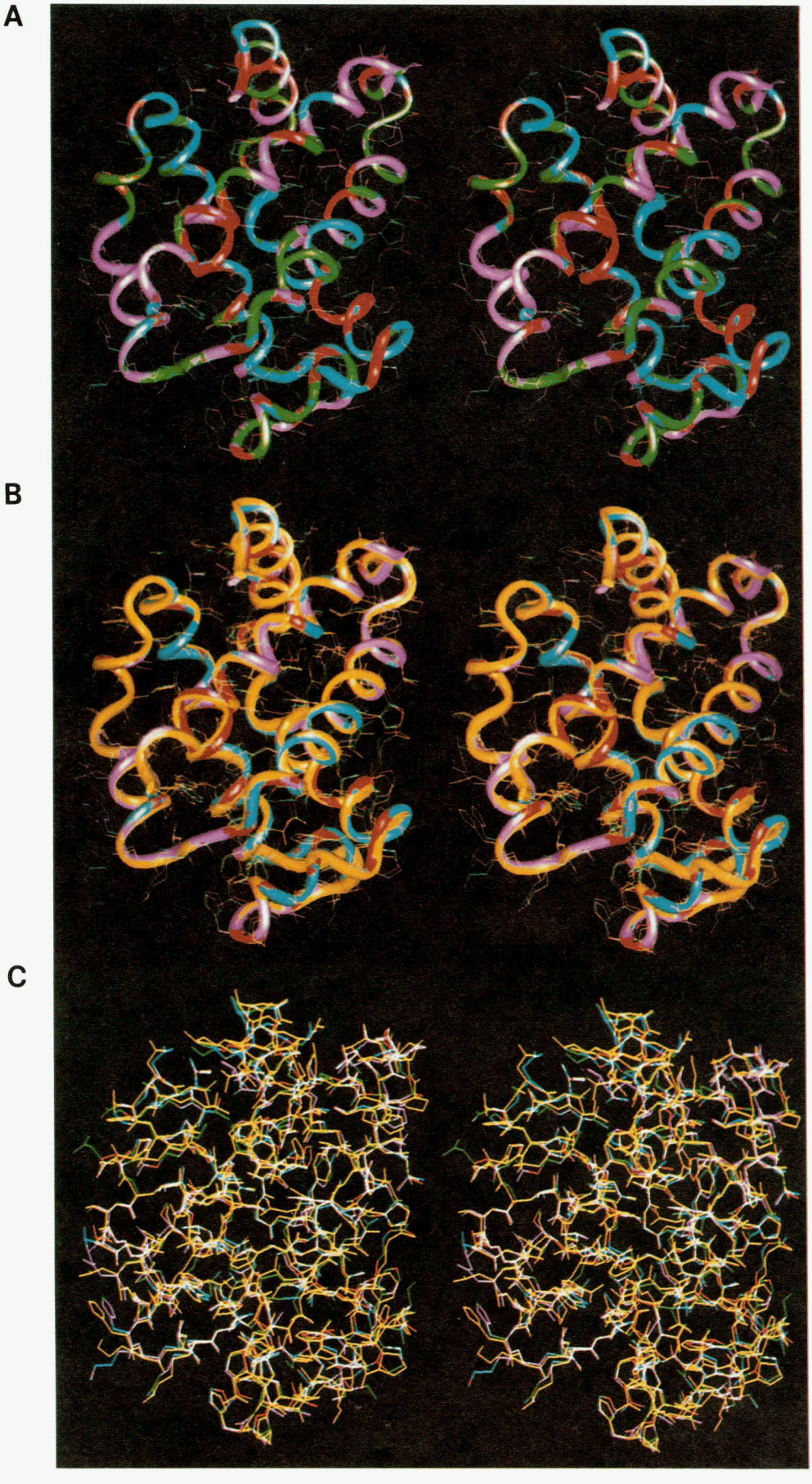


Fig. 1. A: Stereo view of the α-chain of the three selected models (2DHB #11 [red], 2DHB #3 [pink], and 2DHB #13 [cyan]) of horse hemoglobin superimposed on the crystal structure of α-chain of human (green) hemoglobin. **B:** Stereo view of the α-chain of the three selected models (cyan, pink, and red) of horse hemoglobin superimposed on the crystal structure of α-chain of horse (yellow) hemoglobin. **C:** Stereo view of the α-chain of the three selected models (cyan, pink, and red) of horse hemoglobin superimposed on the crystal structures of α-chain of human (green) and horse (yellow) hemoglobins.

Table 1. *Comparison of root mean square (rms) deviations of the models of the α-chain of horse hemoglobin from the crystal structures of α-chains of both the human (PDB code: 2HHB) and horse (PDB code: 2DHB) hemoglobins*

Protein 1	Protein 2	Trace	Backbone	All atoms
2DHB	2HHB	0.581	0.659	–
2DHB	2DHB_11	0.581	0.659	1.026
2DHB	2DHB_3	0.581	0.659	1.044
2DHB	2DHB_13	0.581	0.659	1.033
2HHB	2DHB_11	0.001	0.001	–
2HHB	2DHB_3	0.001	0.001	–
2HHB	2DHB_13	0.001	0.001	–

structures are consistent with the template constraints. The rms deviation of the model structures with the crystal structure of horse hemoglobin are of the same order as that of the crystal structures of horse (2.8 Å resolution) and human (1.74 Å resolution) hemoglobin α-chains. The rms deviation of all the atoms of the model structures with the horse hemoglobin crystal structure is 1.0 Å.

Having established that the model structures are consistent with the template constraints, we decided to test our method for modeling proteins with a lower percentage of sequence identity in the core region. For this purpose, we created a hypothetical worst-case template of a protein domain from the PDB by replacing the interior residues with alanines and reconstructing the interior side chains of the protein using our methodology. This represents a case where the model protein bears the least sequence identity with the interior side chains of the template protein, thus testing the efficiency of this method in packing the interior of the protein. We utilized the variable light chain domain of the Fab fragment (VL3FAB) from a crystal structure in the PDB (PDB code: 3FAB). The ribbon diagram of the tertiary fold of this structure is shown in Figure 2A (green). The residues L4, V10, V18, I20, C22, V32, W34, F61, V63, K65, L72, I74, Y85, C87, S89, V98, T103, and L105 forming the interior of this domain were replaced by alanines, and the resulting structure was used as the template. Twenty structures were simulated using the automated programs presented in the Methods and three structures with the least distance errors were selected. These three structures were energy optimized and are shown superimposed on the crystal structure in Figure 2A. The rms deviation of the coordinates held rigid by the template constraints in the three structures are given in Table 2 along with their initial and final energies. The initial potential energies of all the structures are comparable with the crystal structure energies, showing that the simulated structures have no serious short contacts. For the interior residues listed above, χ_1 values in the three simulated structures are compared with those from the crystal structure in Table 2. Whereas similar χ_1 values among the simulated structures reflect unique side chain conformers within the same general packing configuration, dissimilar χ_1 values suggest alter-

Table 2. *Comparison of χ_1 values (degrees) of the interior residues, initial and final energies (kcal), and rms deviations (Å) of three selected simulated structures with the crystal structure of the light chain constant domain of immunoglobulin (PDB code: 3FAB) as described in the text*

Residue	Crystal χ_1	VL3FAB #4 χ_1	VL3FAB #20 χ_1	VL3FAB #5 χ_1
L4	−77	−149	−70	−155
V10	−180	−77	−151	−80
V18	−70	−62	−166	−160
I20	−57	−40	−69	94
C22	62	−158	−161	180
V32	75	71	−165	60
W34	−81	−65	−103	−92
F61	−68	−52	−49	−43
V63	69	65	75	84
K65	−169	−154	−74	−76
L72	−164	−157	−161	−165
I74	174	49	−180	−61
Y85	−66	−131	−71	−86
C87	66	−170	−168	−165
S89	−171	66	59	−162
V98	−72	−171	−93	−60
T103	−55	44	63	174
L105	−152	−71	−124	43
Initial energy	2,477	2,702	2,714	2,723
Final energy	195	223	199	204
rms deviation	–	0.63	0.56	0.66

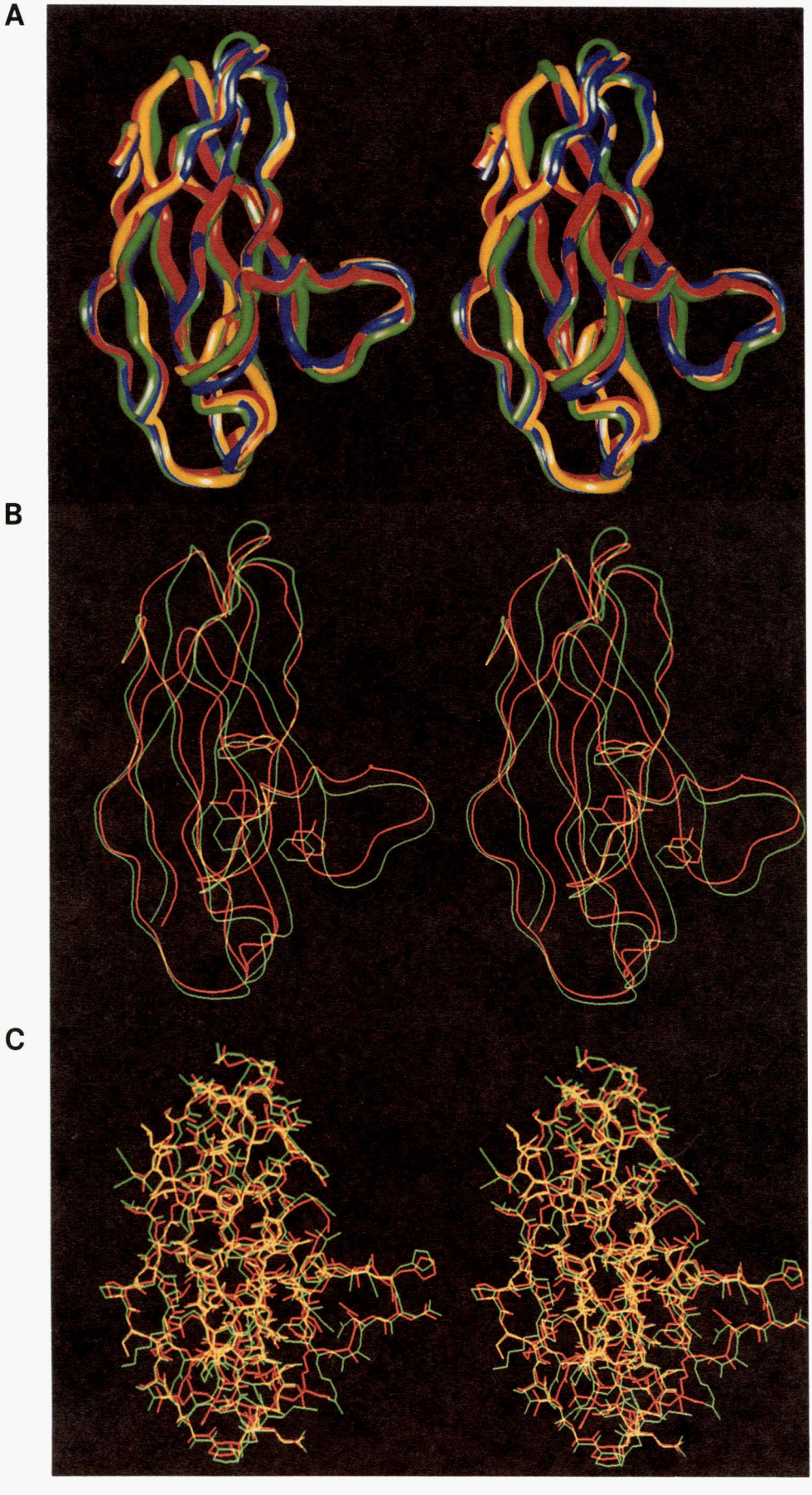


Fig. 2. A: Stereo view of ribbon diagram of the light chain constant domain of immunoglobulin (PDB code: 3FAB) crystal structure (green) superimposed on the ribbon diagram of the three selected structures: VL3FAB #4, VL3FAB #5, and VL3FAB #20. **B:** Stereo view of ribbon diagram of VL3FAB #4 (red) superimposed on the crystal structure (green). W38, F58, and Y82 residues of VL3FAB #4 (red), VL3FAB #5 (yellow), and VL3FAB #20 (pink) are shown relative to these residues in crystal structure (green). **C:** Stereo view of VL3FAB #4 (red) superimposed on the crystal structure (green).

```
MURIL4:   -----HGCDKNHLREIIGILNEVTGEGTPCTEMDVPNVLTATKNT
HUMIL4:   -EAEAHKCDIT-LQEIIKTLNSLTEQKTLCTELTVTDIFAASKDT
TEMPLT:   ------------LQEIIKTLNSLTEQKTLCTELTVTDIFAASKDT

MURIL4:   TESELVCRASKVLRIFYLKHGK-TPCLKKNS-------SVLMELQ
HUMIL4:   TEKETFCRAATVLRQFYSHHEKDTRCLGATAQQFHRHKQLIRFLK
TEMPLT:   TEKETFCRAATVLRQFYSHHEK---CL------------LIRFLK

MURIL4:   RLFRAFRCLDSSISCTMNESKSTSLKDFLESLKSIMQMDYS----
HUMIL4:   RLDRNLWGLAGLNSCPVKEADQSTLENFLERLKTIMREKYSKCSS
TEMPLT:   RLDRNLWGLAGLNSCPVKEADQSTLENFLERLKTIMREKYSKCSS
```

Fig. 3. Sequence alignment between murine interleukin (IL)-4 and human IL-4 used in constructing the model of murine IL-4. TEMPLT sequence shows residues of human IL-4 used as template in building the model. Residues are color coded as follows: lime, residues involved in turns in backbone; blue, basic; red, acidic; green, hydrophobic; yellow, cysteine; black, all others.

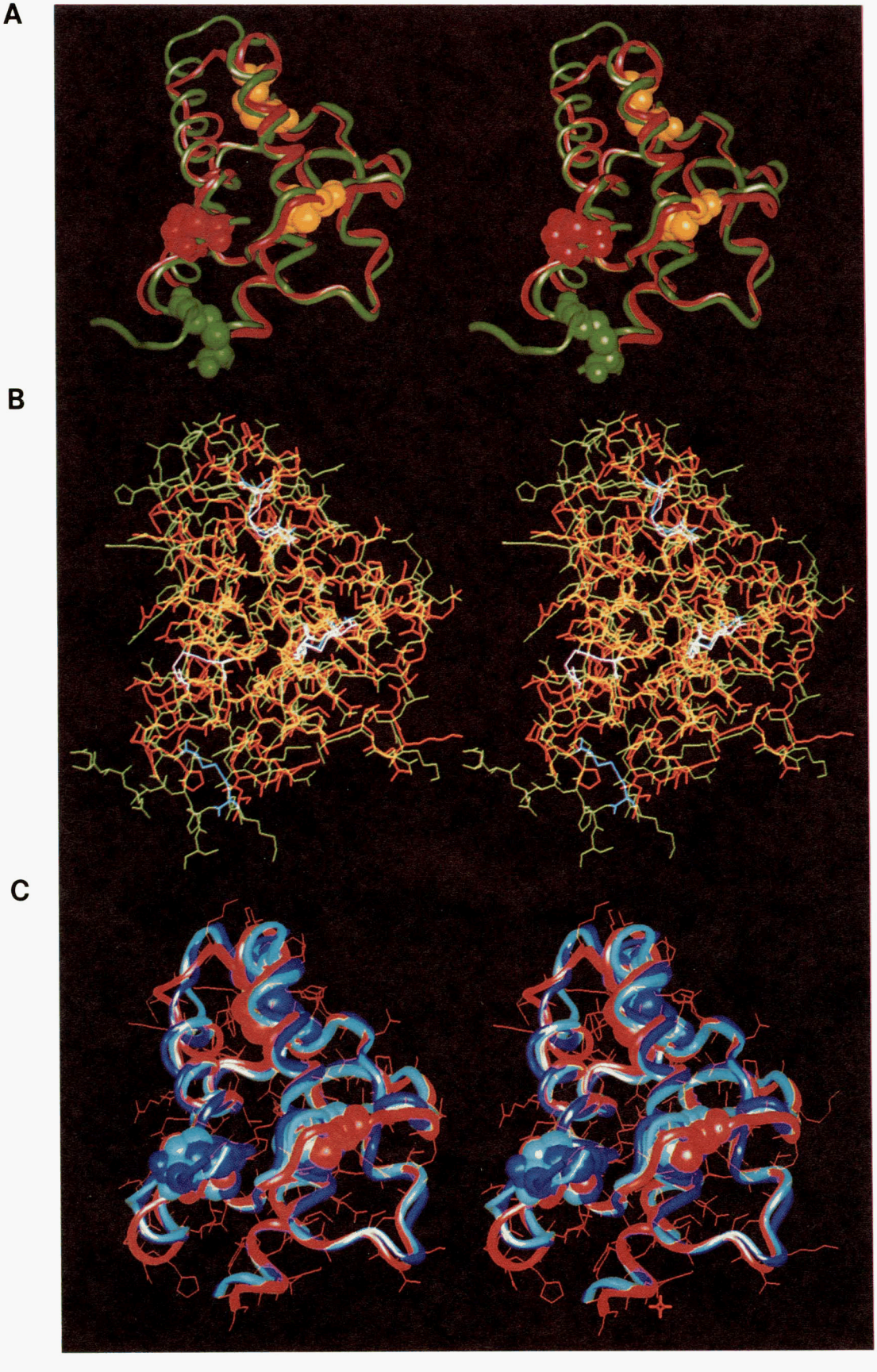
nate packing arrangements for the interior of the protein, which satisfy the distance constraints. It is interesting to note that one of the simulated structures, VL3FAB #4, shown superimposed on the crystal structure in Figure 2B, has W34 and F61 oriented in the same fashion as in the crystal structure. Structure VL3FAB #20 has 11 out of 18 interior side chains, including all ring side chains, with χ_1 values comparable to the crystal structure. In all three selected structures the three ring side chains are found to occupy the same volume element as in the crystal structure (Fig. 2B).

The second test case establishes that our method can preserve structure conservation constraints and results in well-packed structures with no short contacts and no bond deformations even in the cases with weak sequence identity in the core region. This is also reinforced by the similarities in the potential energies of the model structures, which are comparable to the crystal structure. Thus, all the simulated structures provide a good starting structure for energy-based refinements without much user interaction. In this test case, an average of 64% of the ring side chains assumed the same orientation as in the crystal structure, suggesting the important role of interior packing in protein folding.

To test our method for generating structures for proteins with an overall lower percentage of identity, large insertion/deletion regions and nonhomologous disulfides, the structure of murine IL-4 was simulated using the 3D solution structure of human IL-4 (Powers et al., 1992; Smith et al., 1992) as the template. Figure 3 shows the sequence alignment of the two proteins used in building the model. There are two deletion sites in murine IL-4, and they are positioned such that the regions of strong sequence similarity are minimally perturbed. In Figure 3, the sequence TEMPLT consists of all amino acids from the human IL-4 sequence except those near insertion/deletion regions. A minimum of two residues from either side of the insertion/deletion sites are removed in the TEMPLT structure to allow for the required flexibility in simulating the conformations of the deletion regions.

The alternate disulfide link in murine IL-4 between Cys-3 and Cys-85, as compared to human IL-4 (Carr et al., 1991), was accommodated with the addition of disulfide constraints during the simulation. As we noted previously (Curtis et al., 1991), the third helix bears the least similarity in the sequence alignment between murine and human IL-4, and, without additional information, the placement of the large deletion site is ambiguous. However, to accommodate the nonhomologous disulfide bridge found in murine IL-4 between Cys-3 and Cys-85, this deletion was placed near the beginning of the third helix. The upper and lower bounds for the atoms in the cysteines involved in the disulfide are taken from Sowdhamini et al. (1989). Three structures were simulated and energy minimized. In Figure 4A (ribbon) and 4B (atomic detailed), one of the model structures of murine IL-4 (red) is shown superimposed on the human IL-4 structure (green). The two conserved disulfide bridges between human and murine IL-4 are shown in yellow in Figure 4A, whereas the nonhomologous disulfides are rendered in the respective colors of the protein structures. The rms deviation of the backbone atoms between these two structures is 0.64 Å when superimposed using the TEMPLT residues. Figure 4C shows three structures of murine IL-4 embedded using our distance geometry method. The ensemble of conformations for the variable regions in the models have no short contacts and retain trans configurations. The loop conformations among the three struc-

Fig. 4. A: Stereo view of ribbon representations of murine IL-4 (red) model superposed on the NMR structure of human IL-4 (green). Yellow CPK shows the two homologous disulfides in both the structures. Green CPK shows the nonhomologous disulfide in human IL-4, and red CPK shows the nonhomologous disulfide in murine IL-4. **B:** Stereo view of atomic detailed structures of murine IL-4 (red) model superposed on the NMR structure of human IL-4 (green). The three disulfides in human IL-4 structure are shown in cyan and the three disulfides of murine IL-4 are shown in pink. **C:** Stereo view of ribbon representations of three simulated structures of murine IL-4 showing the different loop conformations. Disulfide cross-links in the three structures are shown in CPK. Full atomic structure in red is shown for one simulated structure.


A
B
C

tures generated by this method vary considerably and reflect the general problem related to modeling loops in proteins. However, the short portion of β-sheet in the two overhand loop regions shown in the solution structures of human IL-4 and the helical turns shown in one of the structures (Powers et al., 1992) are conserved in the murine model. The three disulfides are classified as grade B disulfides in the murine IL-4 model (terminology of Sowdhamini et al., 1989). It should be mentioned that in the NMR structure of human IL-4 used as the template, the three disulfides are of grades E (C7–C131), B (C28–C69), and C (C50–C103). The simulated geometry of the disulfide crosslink between Cys-3 and Cys-85 in the murine IL-4 model, for which there is no homologous disulfide in the human IL-4 structure, shows that the distance constraints used to simulate disulfide crosslinks are sufficient to generate converged structures. All structures reported here are minimized to an rms of 0.1 kcal/Å, and minimization is done only with nonhydrogen atoms, due to the fact that only nonhydrogen atoms are simulated. Starting with the sequence alignment file shown in Figure 3, the CPU time necessary to simulate one structure of murine IL-4 is roughly 3 h on a Model 4D/220 Silicon Graphics workstation.

Discussion

The method described here has inherent advantages over conventional homology modeling methods. In our method, there are no short contacts in the simulated structures that usually result in unacceptable strains during the energy refinement step. In conventional methods, the bond stretching and the bond bending strains are kept negligible at the expense of short contacts. Often, these contacts cannot be eliminated by minimization procedures and are generally rectified interactively by rotating single bonds of amino acids involved in short contacts. Rotations around main-chain bonds required to remove short contacts often are not feasible solutions to this modeling problem. For example, in larger proteins, individual bond rotations are time consuming and generally not practical. It should be mentioned that although energy refinement procedures are very efficient in removing strains in bond stretching and bending forces, these methods may actually introduce additional strain while removing short contacts.

Apart from being laborious, conventional methods do not accommodate the cooperative movements in the atomic positions, especially in residues that comprise the interior of a protein. In contrast, we demonstrate that the structures embedded by using our method are optimized by forcing the distance between any two atoms within the lower and upper bounds, thus avoiding close contacts. An additional advantage of the distance geometry method is the random embedding of different structures, allowing all possible positions for side-chain atoms to be explored, thus filling the interior of a protein uniformly. Furthermore, in the distance geometry approach, all the side-chain atoms are treated with the same weighting, leading to the optimal packing of interior residues even when the sequence similarity between the residues in the template and model protein is weak in structurally conserved regions. Our method also enables simulation of structures with alternative packing of the interior residues. This property should be very useful in choosing single-site mutations for structure–function studies of proteins where alternative interior packings may lead to profound effects on exterior conformations.

Our method expands on the earlier works of Srinivasan et al. (1986) and Havel and Snow (1991). We demonstrate that sequences with large insertions and deletions that contain nonhomologous disulfide placements relative to the template structure can still be effectively simulated when additional constraints are utilized. Apart from modeling proteins, the distance constraints used for simulating disulfide crosslinks could be used during the solution of structure sets from NMR-derived constraints. In addition, we provide a totally automated method, once the alignment file is generated, for preparing the constraints file and yielding simulated structures in a reasonable time frame when using the processing power available with desktop workstations.

Our ability to simulate the structure of murine IL-4 provides some insight into the structure–function aspects of four helix bundle cytokines that bind to a homologous family of cell surface receptors, the hematopoietin receptor gene superfamily (Cosman et al., 1990). Diversity in disulfide topology is found among the structurally homologous helical cytokines (Bazan, 1990). We postulate that in helical cytokines, the disulfides play an important role in retaining the functionally active conformations necessary for receptor binding. In IL-4, there is interspecies diversity in disulfide topology between the murine and human proteins (Carr et al., 1991). In our model, the nonhomologous disulfide causes conformational differences between the two species near the C-terminal of the third and fourth helices. In the ligand–receptor complex crystal structure of growth hormone, a four helix bundle with very similar topology to IL-4, and its receptor, a member of the hematopoietin receptor gene superfamily, the C-terminal ends of the third and fourth helices are shown to form the high and low affinity receptor binding epitopes (de Vos et al., 1992). This leads to the postulation that the structural differences between murine and human IL-4 caused by the nonhomologous disulfide are responsible for the lack of cross-species competition in receptor binding. Based on our demonstrated ability to model nonhomologous disulfide pairings, we plan to use this method to model helical cytokines using experimentally determined disulfide pairings to supply additional constraints.

The other site of major structural difference between murine and human IL-4 is found near the beginning of

the third helix caused by the large deletion in the murine IL-4 sequence. Our method provides a means to model this long deletion a priori, without the need to use a knowledge-based approach. This region is also found to form a 3_{10}-helix in the crystal structure of granulocyte-macrophage colony-stimulating factor (GM-CSF) (Diederichs et al., 1991; Walter et al., 1992), suggesting a general feature of conformational diversity near the N-terminus of the third helix among helical cytokines.

The rms deviation of the backbone atoms in the structurally conserved regions of the simulated structure of murine IL-4 is lower than might be expected. In our murine IL-4 model, the rms deviation is 0.64 Å, compared to an expected value of ~1.1 Å based on the percentage of identical residues in the "common core" as defined by Chothia and Lesk (1986). However, analysis of the true core region as defined by the NMR structure for human IL-4 (Powers et al., 1992) reveals 79% identity between the human and murine sequences. Our value of 0.64 Å is consistent with a core identity of 79% as observed by Chothia and Lesk (1986) due to the fact that the difference in rms deviation is greatly affected between the values of 50–80% sequence identity in the core region. If, in fact, our method yields structures with a lower rms deviation than would be expected, this is due to the template atoms being held rigid during the embedding process. A future implementation of our method will allow for consideration of the percentage identity in the structurally conserved regions by incorporating the ability for template atoms to have limited flexibility during structure simulation.

In addition to homology modeling, future applications of our method may include the automated generation of structures from protein sequence databases, simulation of structures with multiple chains, modeling protein structures de novo without the benefit of a homologous template, and the use of multiple templates in modeling and in the solution of crystal structures using the molecular replacement method. Examples of these applications are highlighted below.

Recently, an approach to solving the inverse protein folding problem has been proposed by Bowie et al. (1991). In this approach, the statistical preference of an amino acid to reside in a given environment is matched with the environment of every position in a given template. This method is shown to be successful in generating candidates for structurally similar proteins from protein sequence databases. The growing number of known 3D structures in the PDB offers a large repertoire to model the structures of proteins identified by this novel approach. Thus, our automated method could utilize the Bowie et al. method to identify candidate proteins, which would then be automatically "folded" based on a known template.

The proteins in the intercrine family form different quaternary structures as shown in the 3D structures of IL-8 (Clore et al., 1990) and platelet factor 4 (Charles et al., 1989). In distance space, chemical constraints between the chains can be removed during the simultaneous simulation of the monomeric units and then added back as distance constraints to simulate quaternary structure. It should be noted that by using conventional methods, which use internal coordinates to build homology models, simulation of multiple chain structures could be impractical because valuable distance data regarding monomer orientation in quaternary structure are unavailable. Other inherent advantages of using distance geometry over the conventional methods include modeling proteins with no known homologous structures and simulating an ensemble of topologies consistent with experimentally determined disulfide and chemical constraints. Additionally, structure-function data should be useful in the distance geometry method, providing information that may be translated into additional distance constraints. For example, spectroscopy data may indicate a hydrogen bond between aromatic and acidic side chains. These data have specific implications in distance space that can be easily factored into the model. Our method should provide useful modeling data to suggest additional experiments for structure-function studies, thereby allowing an interactive process for homology modeling.

Multiple templates could be used to model a protein by constructing a chimeric template structure. This could be achieved by aligning the sequences of all the template proteins to the model protein highlighting the regions bearing high sequence identities. The application of this technique would be useful for the construction of highly precise models of proteins where many members are in a family with high levels of sequence identity and where several structures have been solved. This approach would have a limited utility in traditional applications of homology modeling where the sequence identity may be relatively low and only one or two template structures are known. Thus, we will address the use of multiple templates in a subsequent implementation of our method.

Models of proteins built using template structures are also used in protein crystallography. Such models can be simulated using our method and could be used as starting structures in solving the structure of a protein using the molecular replacement method. For example, the low rms deviation between the horse hemoglobin crystal structure and the simulated structures suggests that the murine IL-4 models could be used to solve the crystal structure of murine IL-4 with a native data set.

Although our automated distance geometry method solves many of the problems associated with homology modeling, some basic problems remain. For example, even though our observations show that simulated structures are usually energetically comparable to crystal structures, there is no simple rule to select the best structure from the ensemble embedded by distance geometry. This will remain a problem until the correlation of potential energy functions with the biological activity encoded by a

protein's 3D structure in its native environment is more completely understood. Thus, experimentally derived structure–function data will remain a key element to picking the best model, providing a healthy impetus for interdependence between computational and protein chemistry.

Methods

Starting from a sequence alignment between the model and template proteins, structures of the model protein are simulated using an automated software package, FOLDER. The various steps in the procedure are described below. Each step is accomplished by automated programs that are chained together resulting in the necessary input files for embedding structures using DGEOM. Once these files are created, DGEOM can be run on any available hardware platform, including supercomputers, enabling the simulation of much larger proteins, such as receptors, in a reasonable time frame.

Step 1

A helical conformation is assigned to every amino acid in the protein to be modeled, and Cartesian coordinates are generated for every atom in the model protein using the algorithm of Sundaram and Srinivasan (1979), which has been implemented by the program BUILDER. This helical model is used only to compute the chemical constraints between the atoms in the model protein for which no template constraints would apply. The arbitrary assignment of helical conformation to the peptide backbone and the choice of side-chain rotamers does not impose any restrictions on subsequent stages of the modeling procedure.

Step 2

The sequence alignment between the template and model proteins is used to create a hybrid coordinate set for the model protein. This is done by using a matrix that assigns positional identity among the side-chain atoms in various amino acids (Table 3). In this table, the columns represent the branch location and relative position of the atoms in a side chain in accordance with IUPAC-IUB nomenclature (1970). For every atom in the amino acid side chain, an entry of 1 is made in the respective column. Thus, for identical residues in model and template sequences, all the coordinates in the model structure are replaced by the corresponding template coordinates. For nonidentical residues, an atom-by-atom match is obtained from the matrix shown in Table 3, and the coordinates of each model atom are updated if a match in the template is found. For example, if alanine is replaced by arginine in the model, the coordinates of the arginine atoms N, Cα, C′, O, and Cβ in the model are updated by the template alanine coordinates. To accommodate insertions, the coordinates for all atoms in the insertion residues are retained from the model. The resulting structure is a hybrid structure with coordinates from two reference frames shown in Figure 5A,B. The hybrid structure has artificially long bonds arising from mixing coordinates from two reference frames without any transformations. In Figure 5A, the peptide backbone of human IL-4 (green and yellow) and the helical model of murine IL-4 (blue and magenta) are shown. The two sets of long bonds correspond to the two regions of insertion and deletion in murine IL-4. The number of long bonds in Figure 5A,B represents the atomic coordinates for the atoms in murine IL-4 model that have no homologous atoms in the template. In Figure 5C, the hybrid structure (red) is shown superimposed on the human IL-4 template (green) and the murine IL-4 model (blue). This step uses the program HOMOLOGY.

Step 3

Atoms in the hybrid coordinate file that are assigned coordinates from the template protein are grouped together to form a rigid molecule, and the atoms that are assigned coordinates from the helical model in Step 1 are grouped as a flexible molecule. A list of geometrical constraints is generated by using the program CONSTRAINT to ensure

Table 3. *Positional identity matrix for side-chain atoms of the amino acids used in constructing the hybrid model from a template*[a]

	B (5)	G1 (6)	G2 (7)	D1 (8)	D2 (9)	E1 (10)	E2 (11)	E3 (12)	Z1 (13)	Z2 (14)	Z3 (15)	H1 (16)	H2 (17)
Ala	1	0	0	0	0	0	0	0	0	0	0	0	0
Arg	1	1	0	1	0	1	0	0	1	0	0	1	1
Asn	1	1	0	1	1	0	0	0	0	0	0	0	0
Asp	1	1	0	1	1	0	0	0	0	0	0	0	0
Cys	1	1	0	0	0	0	0	0	0	0	0	0	0
Glu	1	1	0	1	0	1	1	0	0	0	0	0	0
Gln	1	1	0	1	0	1	1	0	0	0	0	0	0
Gly	0	0	0	0	0	0	0	0	0	0	0	0	0
His	1	1	0	1	1	1	1	0	0	0	0	0	0
Ile	1	1	1	1	0	0	0	0	0	0	0	0	0
Leu	1	1	0	1	1	0	0	0	0	0	0	0	0
Lys	1	1	0	1	0	1	0	0	1	0	0	0	0
Met	1	1	0	1	0	1	0	0	0	0	0	0	0
Phe	1	1	0	1	1	1	1	0	1	0	0	0	0
Pro	1	1	0	1	0	0	0	0	0	0	0	0	0
Ser	1	1	0	0	0	0	0	0	0	0	0	0	0
Thr	1	1	1	0	0	0	0	0	0	0	0	0	0
Trp	1	1	0	1	1	1	1	1	0	1	1	1	0
Tyr	1	1	0	1	1	1	1	0	1	0	0	1	0
Val	1	1	1	0	0	0	0	0	0	0	0	0	0

[a] Rows represent amino acids and columns represent positional identity as defined in the Methods. Homologous atoms in different amino acids are aligned in columns with an entry of 1. Positions in rows representing the unfilled position in an amino acid are given an entry of 0.

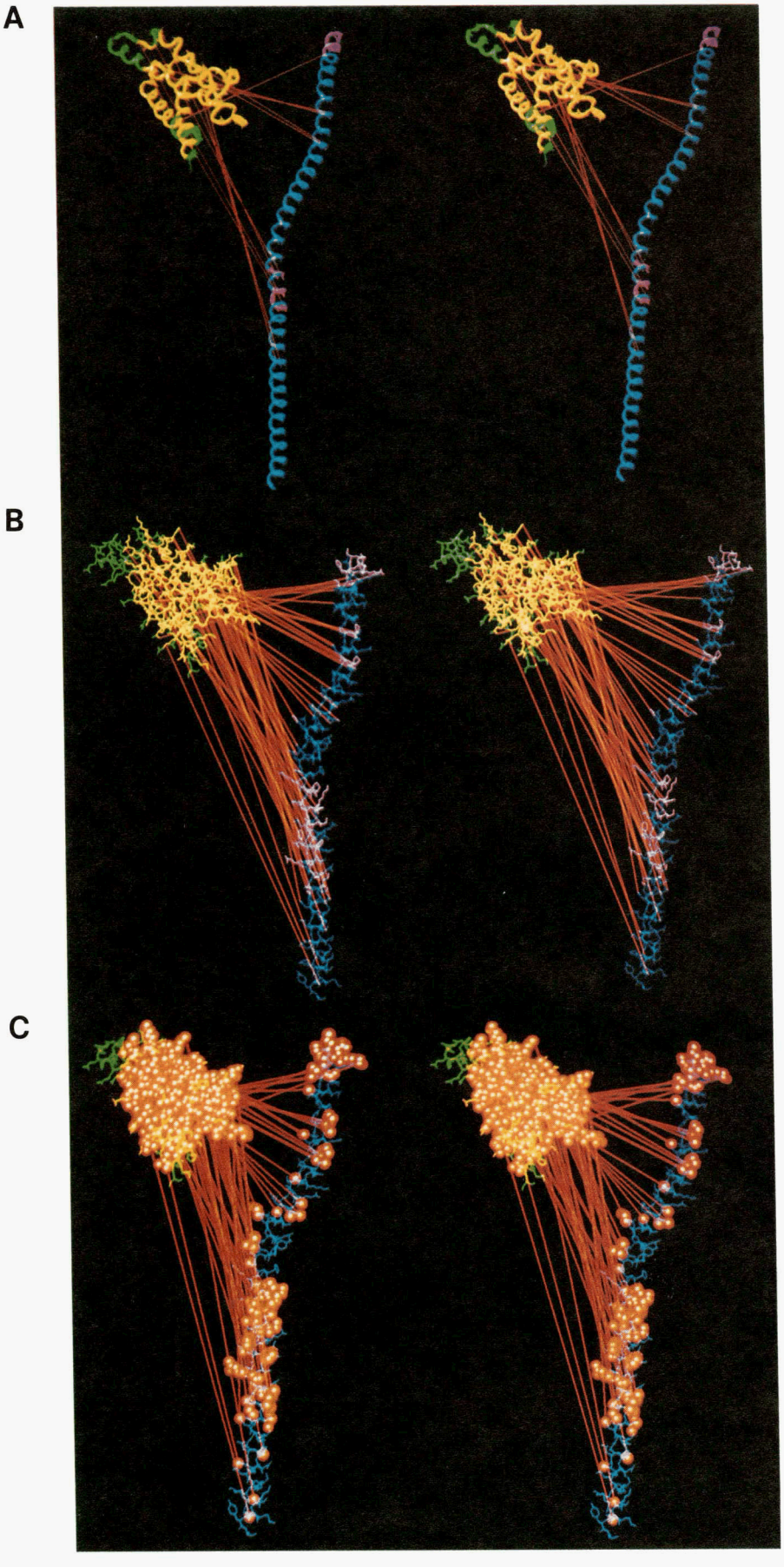


Fig. 5. A: Stereo view of the backbone atoms of the hybrid structure of murine IL-4 constructed by mixing coordinates from the human IL-4 (green) and helical model of murine IL-4 (blue) as described in the Methods, Step 2. In the hybrid structure (red), atoms taken from the template structure are seen yellow because of color blending red with green; coordinates of atoms taken from the helical model are seen in magenta from color blending red with blue; the long red lines are chemical bonds connecting bonded atoms in the two reference frames. **B:** All atoms representation of hybrid structure, template structure, and helical model. Color scheme is the same as in A. **C:** CPK representation in red shows the atoms of the hybrid structure created from template (green) and model (blue) structures.

standard geometry when replacing the long bonds between the chemically bonded atoms in the hybrid coordinate file. In addition to the above constraints, we also impose a fit of the trans conformation of all peptide bonds, chirality constraints at Cβ atoms of isoleucine and threonine, and geometrical constraints between the cysteines forming disulfide crosslinks. CONSTRAINT has been optimized for various amino acid substitutions with varied side-chain geometries.

The above three steps are chained together in the package FOLDER. The sequence alignment file shown in Figure 3 and the template structure file are used as inputs to this automated package to create the two necessary input files for DGEOM. The two files created by the package are the hybrid coordinate file for the model protein and the distance constraints file. DGEOM creates interatomic distance constraints for all the atoms in the hybrid structure consistent with the rigidity constraints of the first molecule. DGEOM also imposes chirality constraints at Cα atoms. This distance matrix is then updated with the distance constraint file created by CONSTRAINT. DGEOM then computes an interatomic distance matrix for every atom pair in the hybrid structure. Interatomic distance bounds for atoms in the rigid molecule are set equal to the distance found in the hybrid structure. The lower distance bounds for all other nonbonded atom pairs (between rigid and flexible or between flexible and flexible hybrid molecules) are set equal to the sum of their van der Waals radii. The upper distance bounds are set equal to the dimensions of the template protein. Using all of the input distance constraints derived from the FOLDER package, DGEOM embeds an ensemble of structures after smoothing the upper and lower bounds.

The structures are then energy optimized using the program BIOGRAF (Version 2.1, Molecular Simulations, Inc.). However, any method of minimization should prove useful due to the excellent starting structures generated by the distance geometry method.

References

Bazan, J.F. (1990). Haemopoietic receptors and helical cytokines. *Immunol. Today 11*, 350–354.

Blaney, J.M., Crippen, G.M., Dearing, A., & Dixon, J.S. (1989). *Quantum Chemistry Program Exchange, Program #590*. Indiana University, Bloomington, Indiana.

Bowie, J.U., Luthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*, 164–170.

Browne, W.J., North, A.C.T., Phillips, D.C., Brew, K., Vanaman, T.C., & Hill, R.L. (1969). A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. *J. Mol. Biol. 42*, 65–86.

Bruccoleri, R.E. & Karplus, M. (1987). Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers 26*, 137–168.

Carr, C., Aykent, S., Kimack, N.M., & Levine, A.D. (1991). Disulfide assignments in recombinant mouse and human interleukin-4. *Biochemistry 30*, 1515–1523

Charles, R.S., Walz, D.A., & Edwards, F.P. (1989). The three-dimensional structure of bovine platelet factor 4 at 3.0-Å resolution. *J. Biol. Chem. 264*, 2092–2099.

Chothia, C. & Lesk, A.M. (1982). Evolution of proteins formed by β-sheets. I. Plastocyanin and azurin. *J. Mol. Biol. 160*, 309–323.

Chothia, C. & Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J. 5*, 823–826.

Chothia, C., Lesk, A.M., Tramantono, A., Levitt, M., Smith-Gill, S.J., Air, G., Sheriff, S., Padlan, E.A., Davies, D., Tulip, W.R., Colman, P.M., Spinelli, S., Alzari, P.M., & Poljak, R.J. (1989). Conformations of immunoglobulin hypervariable regions. *Nature 342*, 877–883.

Claessens, M., Cutsem, E.V., Lasters, I., & Wodak, S. (1989). Modeling the polypeptide backbone with 'spare parts' from known protein structures. *Protein Eng. 2*, 335–345.

Clore, G.M., Appella, E., Yamada, M., Matsushima, K., & Gronenborn, A.M. (1990). Three-dimensional structure of interleukin-8 in solution. *Biochemistry 29*, 1689–1696.

Cosman, D., Lyman, S.D., Idzerda, R.L., Beckmann, M.P., Park, L.S., Goodwin, R.G., & March, C.J. (1990). A new cytokine receptor superfamily. *Trends Biochem. Sci. 215*, 265–270.

Curtis, B.M., Presnell, S.R., Srinivasan, S., Sassenfeld, H., Klinke, R., Jeffery, E., Cosman, D., March, C.J., & Cohen, F.E. (1991). Experimental and theoretical studies of the three-dimensional structure of human interleukin-4. *Proteins Struct. Funct. Genet. 11*, 111–119.

de Vos, A.M., Ultsch, M., & Kossiakoff, A.A. (1992). Human growth hormone and extracellular domain of its receptor: Crystal structure of the complex. *Science 255*, 306–312.

Diederichs, K., Boone, T., & Karplus, P.A. (1991). Novel fold and putative receptor binding site of granulocyte-macrophage colony-stimulating factor. *Science 254*, 1779–1782.

Fine, R.M., Wang, H., Shenkin, P.S., Yarmush, D.L., & Levinthal, C. (1986). Predicting antibody hypervariable loop conformations. II. Minimization and molecular dynamics studies of MOPC603 from many randomly generated loop conformations. *Proteins Struct. Funct. Genet. 1*, 342–362.

Greer, J. (1981). Comparative model building of mammalian serine proteases. *J. Mol. Biol. 153*, 1027–1042.

Havel, T.F. & Snow, M.E. (1991). A new method for building protein conformations from sequence alignments with homologues of known structure. *J. Mol. Biol. 217*, 1–7.

Holmgren, A. & Branden, C.I. (1989). Crystal structure of chaperone protein PapD reveals an immunoglobulin fold. *Nature 342*, 248–251.

IUPAC-IUB Commission on Biochemical Nomenclature. (1970). Abbreviations and symbols for the description of the conformation of polypeptide chains. *J. Biol. Chem. 245*, 6489–6497.

Martin, A.C.R., Cheetham, J.C., & Rees, A.R. (1989). Modeling antibody hypervariable loops: A combined algorithm. *Proc. Natl. Acad. Sci. USA 86*, 9268–9272.

McLachlan, A.D. & Shatton, D.M. (1971). Structural similarities between α-lytic protease of *Myxobacter* 495 and elastase. *Nature 229*, 202–205.

Molecular Simulations, Inc. (1989). *Biograf, Version 2.1*. Molecular Simulations, Inc., Sunnyvale, California.

Moult, J. & James, M.N.G. (1986). An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins Struct. Funct. Genet. 1*, 146–163.

Ponder, J.W. & Richards, F.M. (1987). Tertiary structure templates for proteins: Use of packing criteria in the enumeration of allowed sequences of different structural class. *J. Mol. Biol. 193*, 775–791.

Powers, R., Garrett, D.S., March, C.J., Frieden, E.A., Gronenborn, A.M., & Clore, G.M. (1992). Three-dimensional solution structure of human interleukin-4 by multidimensional heteronuclear magnetic resonance spectroscopy. *Science 256*, 1673–1677.

Ryu, S.-E., Kwong, P.D., Truneh, A., Porter, T.G., Arthos, J., Rosenberg, M., Dai, X., Xuong, N.-H., Axel, R., Sweet, R.W., & Hendrickson, W.A. (1990). Crystal structure of an HIV-binding recombinant fragment of human CD4. *Nature 348*, 419–426.

Smith, L.J., Redfield, C., Boyd, J., Lawrence, G.M.P., Edwards, R.G., Smith, R.A.G., & Dobson, C.M. (1992). Human interleukin 4: The solution structure of a four-helix bundle protein. *J. Mol. Biol. 224*, 899–904.

Snow, M.E. & Amzel, L.M. (1986). Calculating three-dimensional changes in protein structure due to amino-acid substitutions: The variable regions of immunoglobulin. *Proteins 1*, 267–279.

Sowdhamini, R., Srinivasan, N., Shoichet, B., Santi, D.V., Ramakrishnan, C., & Balaram, P. (1989). Stereochemical modeling of disulfide bridges. Criteria for introduction into proteins by site-directed mutagenesis. *Protein Eng. 3*, 95-103.

Srinivasan, S., Shibata, M., & Rein, R. (1986). Multistep modeling of protein structure: Application to bungarotoxin. *Int. J. Quantum Chem. Quantum Biol. Symp. 13*, 167-174.

Sundaram, K. & Srinivasan, S. (1979). Computer simulated modeling of biomolecular systems. *Computer Programs Biomed. 10*, 29-34.

Tramantono, A., Chothia, C., & Lesk, A.M. (1989). Structural determinant of the conformation of medium-sized loops in proteins. *Proteins 6*, 382-394.

Walter, M.R., Cook, W.J., Ealick, S.E., Nagabhushan, T.L., Trotta, P.P., & Bugg, C.E. (1992). Three-dimensional structure of recombinant human granulocyte-macrophage colony-stimulating factor. *J. Mol. Biol. 224*, 1075-1085.

Wang, J., Yan, Y., Garrett, T.P.J., Liu, J., Rodgers, D.W., Garlick, R.L., Tarr, G.E., Husain, Y., Reinherz, E.L., & Harrison, S.C. (1990). Atomic structure of a fragment of human CD4 containing two immunoglobulin-like domains. *Nature 348*, 411-418.