Protein Science (1997), *6*:956–970. Cambridge University Press. Printed in the USA.

Homology modeling using simulated annealing of restrained molecular dynamics and conformational search calculations with CONGEN: Application in predicting the three-dimensional structure of murine homeodomain Msx-1

HAICHENG LI,[1,2] ROBERTO TEJERO,[1,3] DANIEL MONLEON,[3] DONNA BASSOLINO-KLIMAS,[4] CORY ABATE-SHEN,[1,5] ROBERT E. BRUCCOLERI,[4] AND GAETANO T. MONTELIONE[1,2]

[1]Center for Advanced Biotechnology and Medicine, Piscataway, New Jersey 08854-5638
[2]Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, New Jersey 08854-5638
[3]Departamento de Quimica Fisica, Universidad de Valencia, Dr. Moliner 50, 46100-Burjassot, Valencia, Spain
[4]Department of Macromolecular Structure, Bristol-Myers Squibb Pharmaceutical Research Institute, Princeton, New Jersey 08543
[5]Department of Neuroscience and Cell Biology, UMDNJ–Robert Wood Johnson Medical School, Piscataway, New Jersey 08854-5638

(RECEIVED October 15, 1996; ACCEPTED February 4, 1997)

Abstract

We have developed an automatic approach for homology modeling using restrained molecular dynamics and simulated annealing procedures, together with conformational search algorithms available in the molecular mechanics program CONGEN (Bruccoleri RE, Karplus M, 1987, *Biopolymers 26*:137–168). The accuracy of the method is validated by "predicting" structures of two homeodomain proteins with known three-dimensional structures, and then applied to predict the three-dimensional structure of the homeodomain of the murine Msx-1 transcription factor. Regions of the unknown protein structure that are highly homologous to the known template structure are constrained by "homology distance constraints," whereas the conformations of nonhomologous regions of the unknown protein are defined only by the potential energy function. A full energy function (excluding explicit solvent) is employed to ensure that the calculated structures have good conformational energies and are physically reasonable. As in NMR structure determinations, information on the consistency of the structure prediction is obtained by superposition of the resulting family of protein structures. In this paper, our homology modeling algorithm is described and compared with related homology modeling methods using spatial constraints derived from the structures of homologous proteins. The software is then used to predict the DNA-bound structures of three homeodomain proteins from the X-ray crystal structure of the engrailed homeodomain protein (Kissinger CR et al., 1990, *Cell 63*:579–590). The resulting backbone and side-chain conformations of the modeled yeast Matα2 and *D. melanogaster* Antennapedia homeodomains are excellent matches to the corresponding published X-ray crystal (Wolberger C et al., 1991, *Cell 67*:517–528) and NMR (Billeter M et al., 1993, *J Mol Biol 234*:1084–1097) structures, respectively. Examination of these structures of Msx-1 reveals a network of highly conserved surface salt bridges that are proposed to play a role in regulating protein–protein interactions of homeodomains in transcription complexes.

Keywords: conformational energy calculations; conserved surface salt bridges; DNA-binding protein; transcription factor

Reprint requests to: Gaetano T. Montelione, CABM, Rutgers University, 679 Hoes Lane, Piscataway, New Jersey 08854-5638; e-mail: guy@nmrlab.cabm.rutgers.edu.

Abbreviations: Antp, homeodomain of the *D. melanogaster* Antennapedia protein; DG, distance geometry calculations using metric-matrix embedding methods; Conf. E., total conformational energy including electrostatic effects, computed from the CHARMM potential function; Matα2, homeodomain of the yeast Matα2 protein; Msx-1, homeodomain of the murine Msx-1 protein; pdf, probability density function; RMSD, RMS deviation; SARMD, simulated annealing with restrained molecular dynamics; VDW E., van der Waals energy computed from the Lennard–Jones portion of the CHARMM potential function.

The homeodomain is a highly conserved sequence-specific DNA-binding domain that has been found in many transcription factors. First discovered in *Drosophila melanogaster*, homeodomains have been found in almost every organism from nematodes to humans (Kessel & Gruss, 1990; Wang et al., 1993), and have been found to play a fundamental role in directing embryogenesis (Gehring, 1987; Scott et al., 1989; for a recent review see Krumlauf, 1994). The sequence of the homeodomain corresponds to 60 amino acid residues that assemble into three α-helices and one flexible N-terminal arm (Scott et al., 1989; Kissinger et al., 1990; Laughon, 1991; Wolberger, 1996). Structures of several homeodomains in both unliganded and DNA-bound states have been determined by NMR and X-ray crystallography (Otting et al., 1988; Qian et al., 1989, 1993, 1994; Kissinger et al., 1990; Phillips et al., 1991; Wolberger et al., 1991; Assa-Munt et al., 1993; Billeter et al., 1993, 1996; Ceska et al., 1993; Cox et al., 1993; Leiting et al., 1993; Klemm et al., 1994; Sivaraja et al., 1994; Li et al., 1995). These protein structures share a common three-helical chain fold in which the second and third helices are arranged in a helix-turn-helix motif. In the five homeodomain protein–DNA complexes that are available (Kissinger et al., 1990; Wolberger et al., 1991; Billeter et al., 1993; Klemm et al., 1994; Li et al., 1995), the second helix of the helix-turn-helix motif (helix III in Fig. 1) contacts DNA in the major groove and the "N-terminal arm" (residues 1–9 in Fig. 1) contacts DNA in the minor groove.

Homeodomains have a high degree of conservation in primary sequence, tertiary structure, and their mode of interaction with DNA (Gehring et al., 1994; Wolberger, 1996). They all bind to consensus sites that contain a TAAT core motif (Odenwald et al., 1989; Laughon, 1991; Catron et al., 1993; Gehring et al., 1994; Wolberger, 1996). Despite their high conservation, most homeodomain proteins function by binding to specific duplex DNA sequences, and even small differences in their sequence specificity and affinity can have biological significance (Laughon, 1991; Dessain et al., 1992; Ekker et al., 1992; Catron et al., 1993; Kornberg, 1993; Gehring et al., 1994). Recently, homeodomain proteins have also been found to recognize specific RNA molecules (Dubnau & Struhl, 1996; Rivera-Pomar et al., 1996). Thus, understanding the atomic basis of specific homeodomain/nucleic acid interactions is essential to the comprehension of the complex interactions that lead to transcriptional (and perhaps translational) regulation in development.

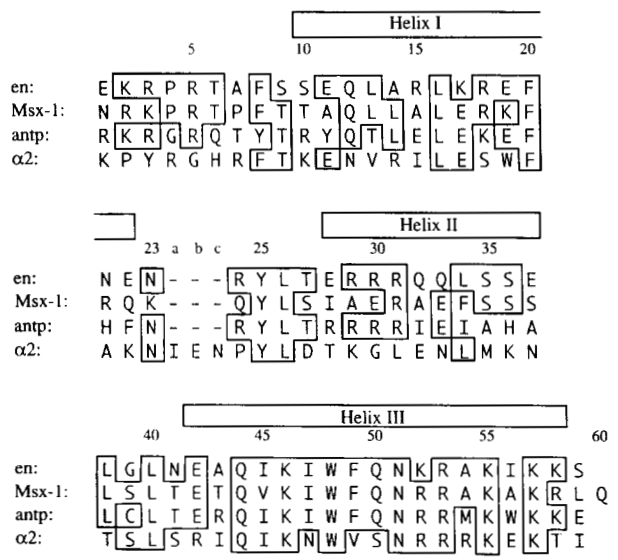

Fig. 1. Amino acid sequence alignments for four homeobox domains: engrailed (en), Msx-1, antennapedia (antp), and Matα2 (α2).

Msx-1 is a homeodomain protein encoded by a member of a relatively small subfamily of homeobox genes expressed in craniofacial structures, the neural tube, and the limbs of the developing embryo (Davidson et al., 1991; Krumlauf, 1994). Expression of Msx-1 in myoblasts inhibits terminal differentiation and induces cell transformation (Song et al., 1992). Msx-1 interacts with DNA sites that contain the consensus sequence (C/G)TAATTG (Catron et al., 1993) and functions as a transcriptional repressor both in vitro and in vivo (Catron et al., 1995, 1996). Like many other homeodomains, Msx-1 appears to regulate cellular proliferation by its ability to repress differentiation-specific target genes.

Because the relative positions of certain structurally and/or functionally crucial atoms should be similar among a family of homologous proteins, the three-dimensional structure of a protein can often be modeled reliably based on the known structures of homologous proteins. Homology modeling has been applied successfully to a number of proteins (see, for example, Greer, 1985, 1990; Chothia et al., 1986; Palmer et al., 1986; Havel & Snow, 1991; Bruccoleri & Novotny, 1992; Bajorath et al., 1993; Brocklehurst & Perham, 1993; Fogolari et al., 1993; Havel, 1993; Srinivasan et al., 1993; for a comprehensive review see Sali, 1995). Such homology models are very useful in certain protein engineering applications that do not require an accurate high-resolution structure, and for accelerating experimental structure determinations by NMR and X-ray crystallography using molecular replacement methods. Several successful approaches for homology modeling (reviewed in Sali, 1995) have included knowledge-based interactive model building (Blundell et al., 1983; Claessens et al., 1989; Bazan, 1990; Bajorath et al., 1993), systematic conformational search (Bruccoleri & Novotny, 1992), combinatorial side-chain conformational analysis (Ponder & Richards, 1987), polypeptide segment matching (Levitt, 1992), conformational threading (Jones et al., 1992), distance geometry and/or simulated annealing calculations using homology constraints (Engh et al., 1990; Havel & Snow, 1991; Fujiyoshi-Yoneda et al., 1991; Brocklehurst & Perham, 1993, Havel, 1993; Snow, 1993; Srinivasan et al., 1993; Sudarsanam et al., 1994), and structure generation by satisfaction of spatial restraints derived from sequence alignments and expressed as probability density functions (Sali & Blundell, 1993). However, there is no general agreement on what is the most reliable method for this process. In this regard, it is important to test methods for homology modeling on available NMR and X-ray crystal structures in order to validate their precision and accuracy. It is also important to develop a database of homology-modeled structures of proteins for which no NMR or X-ray crystal structure is yet available, because later these can be compared with experimentally determined structures to evaluate weaknesses and strengths of different methods.

Many of the existing methods for homology modeling mentioned above result in structures with high-energy overlaps of nonbonded atoms and other unreasonable energetic features; when energy is simply minimized, these structures can become trapped in local energy minima (Scheraga, 1984). Recently, distance geometry (Havel & Snow, 1991; Havel, 1993; Srinivasan et al., 1993; Sudarsanam et al., 1994), simulated annealing with molecular dynamics (Engh et al., 1990; Fujiyoshi-Yoneda et al., 1991; Brocklehurst & Perham, 1993), and variable target function (Sali

& Blundell, 1993) minimization methods like those used for protein structure determination from NMR data have been found to be adopted easily for generating families of homology-modeled structures from "homology constraints." Like NMR structure determinations, the results of these structure generation calculations can be represented as a *family of structures*, each of which satisfies the "homology constraints." Superposition of this family of solutions provides information about the consistency of the homology-modeled structures given the input homology assumptions. Although homology models generated with DG algorithms may sometimes exhibit high energies and biases due to inadequate samplings of solution space, these biases can be eliminated and the conformational energies improved by further refinement of structures using restrained molecular dynamics (Havel, 1993; Sali & Blundell, 1993).

We have developed an automatic and objective approach for homology modeling using simulated annealing with restrained molecular dynamics and conformational search methods available in the molecular mechanics program CONGEN (Bruccoleri & Karplus, 1987, 1990; Bruccoleri, 1993; Bassolino-Klimas et al., 1996). Our hybrid approach (SARMD/search) avoids the metric matrix embedding step used in similar homology modeling methods (Havel & Snow, 1991; Havel, 1993; Srinivasan et al., 1993) or the use of variable-target function minimization (Sali & Blundell, 1993), and generates structures directly from extended starting conformations using restrained molecular dynamics calculations. These structures are further refined using conformational search methods (Bruccoleri & Karplus 1987, 1990; Bruccoleri et al., 1988; Bruccoleri & Novotny, 1992) for improved sampling of low-energy conformations of poorly defined surface loops and side chains. Regions of the unknown protein structure that are highly homologous to the known template structure are constrained by "homology distance constraints," whereas the conformations of nonhomologous regions of the unknown protein are defined only by the potential energy function. A full energy function (excluding explicit solvent) is employed to ensure that the calculated structures have good conformational energies and are physically reasonable. Information on the consistency of the structure prediction is obtained by superposition of the resulting family of protein structures.

In this paper, our homology modeling algorithms are described and used to predict the three-dimensional structures of the yeast homeodomain Matα2 (Scott et al., 1989), the *D. melanogaster* Antennapedia homeodomain (Antp) (Schneuwly et al., 1986), and the murine homeobox domain Msx-1 (Catron et al., 1993) from the X-ray crystal structure of the *D. melanogaster* engrailed homeodomain (Kissinger et al., 1990). The resulting backbone and side-chain conformations of the modeled homeodomains Antp and Matα2 are in excellent agreement with the published NMR (Billeter et al., 1993) and X-ray crystal (Wolberger et al., 1991; Li et al., 1995) structures, respectively. Comparisons of these structures reveal a conserved network of surface salt bridges common to the engrailed, Matα2, Antp, and Msx-1 homeodomains. This family of predicted structures of the Msx-1 homeodomain has been deposited in the Brookhaven Protein Data Bank for future comparisons with NMR and X-ray structures once they are available.

Results

Summaries of our homology modeling algorithms and the protocols used for SARMD/search calculations are presented in Materials and methods.

Helix III of homeodomain Msx-1 from mouse

Our approach for homology modeling by SARMD was first tested by computing the 3D structure of the DNA recognition helix of Msx-1. This polypeptide fragment contains 17 amino acid residues, 11 of which are consensus residues that are conserved in nearly all homeobox domains. Fourteen residues are identical in this segment of Msx-1 and engrailed homeodomains (Fig. 1). A set of 25 structures of Msx-1 helix III was calculated using 323 "homology constraints," and the 10 with the smallest residual constraint violations and lowest conformational energies (as explained in the Materials and methods) were selected for structural analysis. The result of these calculations is a well defined α-helix. A summary of structural statistics is reported in Table 1 and a summary of residual "homology constraint" violations and conformational energies is presented in Table 2. The RMSD of the backbone atoms within the bundle of structures is 0.3 Å. Most side chains are also well-defined in the predicted structure, with RMSDs for all heavy atoms of 0.8 Å within the bundle. Van der Waals energies for the predicted structure are very low, ranging from -4.2 to -4.7 kcal/mol-residue, whereas deviations from ideal bond angles, bond lengths, and planar peptide bonds (Table 1) are within the range of deviations typically seen in high-quality X-ray and NMR structure determinations. These results show that SARMD calculations with CONGEN can be used to generate homology models of this polypeptide segment that exhibit good conformational energies and satisfy the set of homology constraints.

Homeodomain Matα2 from yeast

In order to validate the predictive value of our SARMD/search method, we next homology modeled the three-dimensional structure of the yeast Matα2 homeodomain, whose crystal structure has been determined to 2.7 Å resolution (Wolberger et al., 1991). The X-ray structure of the DNA-bound engrailed homeodomain protein was again used as a template. Matα2 has 27% sequence identity with the engrailed homeodomain. Relative to engrailed, the polypeptide sequence of Matα2 also includes a tripeptide insertion in the polypeptide loop segment between helices I and II (Fig. 1), which is not present in the other homeodomains that we studied. In generating homology models of Matα2 from the structure of engrailed, we assumed that the relative positions of the three helices are well conserved, because the overall homeodomain structure appears to be very well conserved in nature, but that the conformation of this interhelical loop can adjust to accommodate the tripeptide insert. Accordingly, homologous atoms in the Matα2 homeodomain were defined relative to the engrailed template sequence, except in the vicinity of the inserted octapeptide segment of residues 23–27 (Fig. 1). No homologous atoms are defined in this loop region (i.e., for polypeptide segment Asn 23–Ile 23a–Glu 23b–Asn 23c–Pro 24–Tyr 25–Leu 26–Asp 27) because the insertion of three residues into the original five-residue-long sequence would likely change the conformation of the loop. As a result, no constraints are applied in the loop region during the simulated annealing procedure and its conformation is determined solely by energy considerations.

A set of 25 homology-modeled structures of Matα2 were calculated from 919 "homology constraints," and the 10 with the smallest residual constraint violations and lowest conformational energies were selected for structural analysis. This family of 10 structures is shown in Figure 2A. The convergence within these 10 structures is quite good, except in the loop region. Atomic RMSDs

Table 1. *Statistics for homology-modeled structures*

	Msx-1 Helix III[a]	Matα2	Antp	Msx-1
Homology constraints				
Total number of heavy atoms	177	532	555	507
Total number of homologous heavy atoms	142	347	444	375
Total number of homology constraints[b]	323	919	1,000	1,000
Intraresidue [$i = j$]	10	2	4	4
Sequential [$(i - j) = 1$]	37	28	34	30
Longer-range [$(i - j) > 1$]	276	889	962	966
Structural statistics				
Number of final structures	10	10	10	10
RMS constraint violation	0.0067 Å	0.0088 Å	0.0054 Å	0.0053 Å
RMSDs from ideal polypeptide geometries				
Bond angles	1.78°	2.39°	2.23°	2.00°
Bond lengths	0.010 Å	0.010 Å	0.011 Å	0.010 Å
Peptide bond ωs	2.53°	0.66°	0.63°	2.19°
Atomic RMSD[c] to average homology-modeled structure				
Backbone (N, C^α, C′)	0.3 Å[a]	0.5 Å[d]	0.4 Å[e]	0.3 Å[f]
All heavy atoms	0.8 Å[a]	1.2 Å[d]	1.0 Å[e]	0.9 Å[f]
Atomic RMSD[c] to X-ray crystal or average NMR structure				
Backbone (N, C^α, C′)	—	0.9 Å[d]	0.8 Å[e]	—
All heavy atoms	—	2.0 Å[d]	1.7 Å[e]	—
Atomic RMSD[c] of engrailed template structure to X-ray crystal or average NMR structure				
Backbone (N, C^α, C′)	—	0.9 Å[d]	0.8 Å[e]	—

[a]Residues 42–58 of Msx-1.
[b]Each homology constraint corresponds to one upper-bound and one lower-bound distance constraint.
[c]RMSDs are computed only for residues with backbone dihedral order parameters (Hyberts et al., 1992), $S(\phi) + S(\psi) > 1.5$. Residue numbers used here are defined in Figure 1.
[d]For residue range 12–22, 32–60, excluding poorly defined N-terminal, loop, and C-terminal residues.
[e]For residue range 4–57, excluding poorly defined N-terminal and C-terminal residues.
[f]For residue range 3–56, excluding poorly defined N-terminal and C-terminal residues.

for residues 7–58 (excluding the loop region) are 0.5 Å for backbone atoms and 1.2 Å for all heavy atoms (Table 1). Residual constraint violations and conformational energies for this family of homology-modeled Matα2 structures are presented in Table 2. The VDW energies for the 10 calculated structures range from −4.5 to −4.9 kcal/mol-residue, and are similar to values obtained in CONGEN SARMD calculations on the small protein crambin using distance constraints derived from the X-ray crystal structure (Bassolino-Klimas et al., 1996) and for the structures of small proteins using experimental NMR data (Tejero et al., 1996).

The backbone conformation of each homology-modeled Matα2 structure is very similar to that of the X-ray structure (Fig. 2A). Like the engrailed homeodomain template structure, the crystal structure of Matα2 (Wolberger et al., 1991) was determined in complex with DNA. A statistical summary of the comparison between the homology-modeled and X-ray structures of Matα2 is also shown in Table 1. The average RMSD of the backbone conformation (the loop region excluded) between the modeled and the X-ray structures is 0.9 Å. When individual helices are compared, the corresponding RMSD values for backbone atoms range from 0.3 to 0.7 Å. The loop region is not well-defined in the homology models; instead, it is predicted to exist in two families of conformations, one of which includes the backbone conformation observed in the crystal structure (Fig. 2A).

Homeodomain Antennapedia from D. melanogaster

We next modeled a second homeodomain, the Antennapedia homeodomain (Antp) from *D. melanogaster*. As before, we used the engrailed homeodomain crystal structure determined in complex with double-stranded DNA (Kissinger et al., 1990) as the structural template. The sequence identity between the Antp and engrailed homeodomains is 51% (Fig. 1). A family of 25 structures was calculated using 1,000 homology constraints and the 10 structures with lowest constraint violations and conformational energies were selected for structural analysis. A summary of structural statistics for this family is presented in Table 1 and a summary of the residual constraint violations and energies is shown in Table 2. The convergence within the family of 10 calculated structures is very good, with RMSD values of 0.4 Å for the backbone and 1.0 Å for all heavy atoms of residues 4–57. Residual violations and conformational energies for these 10 structures of Antp are summarized in Table 2. The VDW energies of these homology-modeled Antp structures range from −5.2 to −5.7 kcal/mol-residue; these values are typical of good-quality structures in the CONGEN force field (Bassolino-Klimas et al., 1996; Tejero et al., 1996).

These 10 homology-modeled structures were superimposed on the average NMR structure of DNA-bound Antp (Fig. 2B). The backbone conformations of the homology-modeled structures are very similar to the average NMR-determined structure (Table 1),

Table 2. *Summary of residual homology-constraint violations and final energies*

	Numbers of violations and energies for each homology-modeled structure									
Structure no.	1	2	3	4	5	6	7	8	9	10
Msx-1 Helix III[a]										
>0.2 Å	0	0	0	0	0	0	0	0	0	0
0.1–0.2 Å	1	2	0	3	2	4	3	1	4	2
<0.1 Å	13	13	10	12	8	9	12	15	10	11
Total	14	15	10	15	10	13	15	16	14	13
VDW E.[b]	−71.2	−74.2	−77.5	−73.8	−73.7	−78.1	−72.2	−80.1	−77.0	−75.0
Conf. E.[b]	−317.0	−278.9	−287.2	−299.5	−283.2	−300.5	−282.9	−275.0	−285.0	−258.9
Matα2										
>0.2 Å	0	0	0	0	0	0	0	0	0	0
0.1–0.2 Å	2	3	1	0	2	2	2	0	2	1
<0.1 Å	66	65	56	43	70	49	58	40	65	43
Total	68	68	57	43	72	51	60	40	67	44
VDW E.[b]	−285.9	−279.7	−304.7	−297.1	−277.1	−277.4	−289.4	−304.4	−287.9	−292.1
Conf. E.[b]	−1,216.6	−1,269.1	−1,298.6	−1,338.2	−1,156.9	−1,204.9	−1,207.4	−1,335.6	−1,155.3	−1,246.3
Antp										
>0.2 Å	0	0	0	0	0	0	0	0	0	0
0.1–0.2 Å	0	0	0	0	0	1	0	0	0	0
<0.1 Å	50	56	49	64	68	52	52	48	43	58
Total	50	56	49	64	68	53	52	48	43	58
VDW E.[b]	−325.9	−334.2	−323.9	−318.3	−317.2	−322.6	−312.7	−319.3	−314.3	−304.9
Conf. E.[b]	−1,547.5	−1,511.3	−1,516.2	−1,538.2	−1,495.8	−1,526.5	−1,539.9	−1,534.6	−1,543.8	−1,525.5
Msx-1										
>0.2 Å	0	0	0	0	0	0	0	0	0	0
0.1–0.2 Å	0	0	1	0	0	0	0	0	0	0
<0.1 Å	62	53	59	68	65	56	69	60	59	64
Total	62	53	60	68	58	56	69	60	59	64
VDW E.[b]	−292.3	−292.4	−290.0	−295.5	−291.1	−304.9	−280.4	−288.4	−295.6	−292.7
Conf. E.[b]	−1,186.5	−1,223.1	−1,230.4	−1,201.5	−1,215.1	−1,223.6	−1,197.5	−1,281.0	−1,248.7	−1,208.0

[a]Residues 42–58 of Msx-1.
[b]Van der Waal (VDW E.) and conformational (Conf. E.) energies are defined in the text and are reported in units of kcal mol^{-1}.

with backbone RMSD of 0.8 Å for residues 4–57; this value is about the same as the backbone RMSD for residues 4–57 within the family of 16 NMR structures deposited in the Brookhaven Protein Data Bank (Billeter et al., 1993). For each helix, as well as for the helix-turn-helix motif in the structure, backbone RMSD values between the predicted structures and the average NMR structure range from 0.3 to 0.8 Å. For all heavy atoms of the predicted Antp structure, the RMSD for residues 8–56 relative to the average NMR structure is 1.7 Å, slightly higher than the RMSD of 1.3 Å observed for the heavy atoms of the 16 NMR structures relative to the average NMR coordinates. This deviation is understandable because some of the side chains of the NMR structure are constrained by interactions with DNA. In addition, some surface side chains interpreted as single conformers in the NMR analysis may in fact adopt several isoenergetic conformations.

Homeodomain Msx-1 from mouse

From the predictions of atomic coordinates for Matα2 and Antp, it appears that homology modeling using SARMD/search with CONGEN provides a robust and reliable method for predicting homologous homeodomain backbone chain folds. Having validated our approach on two known homeodomain structures, the entire three-dimensional structure of murine homeodomain Msx-1 was then predicted by SARMD/search with CONGEN; Msx-1 has 48% sequence identity with the engrailed homeodomain. Twenty-five conformers were generated, and the 10 with the lowest constraint violations and energies were selected to represent the predicted structure of Msx-1. A superposition of these 10 structures is shown in Figure 3 and statistics for the superpositions of backbone and heavy atoms for these structures are also presented in Table 1. As before, the convergence of calculated structures was quite good; for residues 3–56, the RMSDs between each predicted structure and the average structure are 0.3 Å for backbone atoms and 0.9 Å for all heavy atoms. VDW energies for these homology-modeled structures of Msx-1 range from −4.7 to −5.1 kcal/mol-residue, which are comparable to the corresponding values for Matα2 and Antp homeodomains and typical of good-quality structures in the CONGEN force field (Bassolino-Klimas et al., 1996; Tejero et al., 1996).

Discussion

Homology modeling by SARMD and conformational search using CONGEN

This study provides a benchmark comparison of an automated homology modeling technique using SARMD protocols (Bassolino-Klimas et al., 1996; Tejero et al., 1996) and conformational search

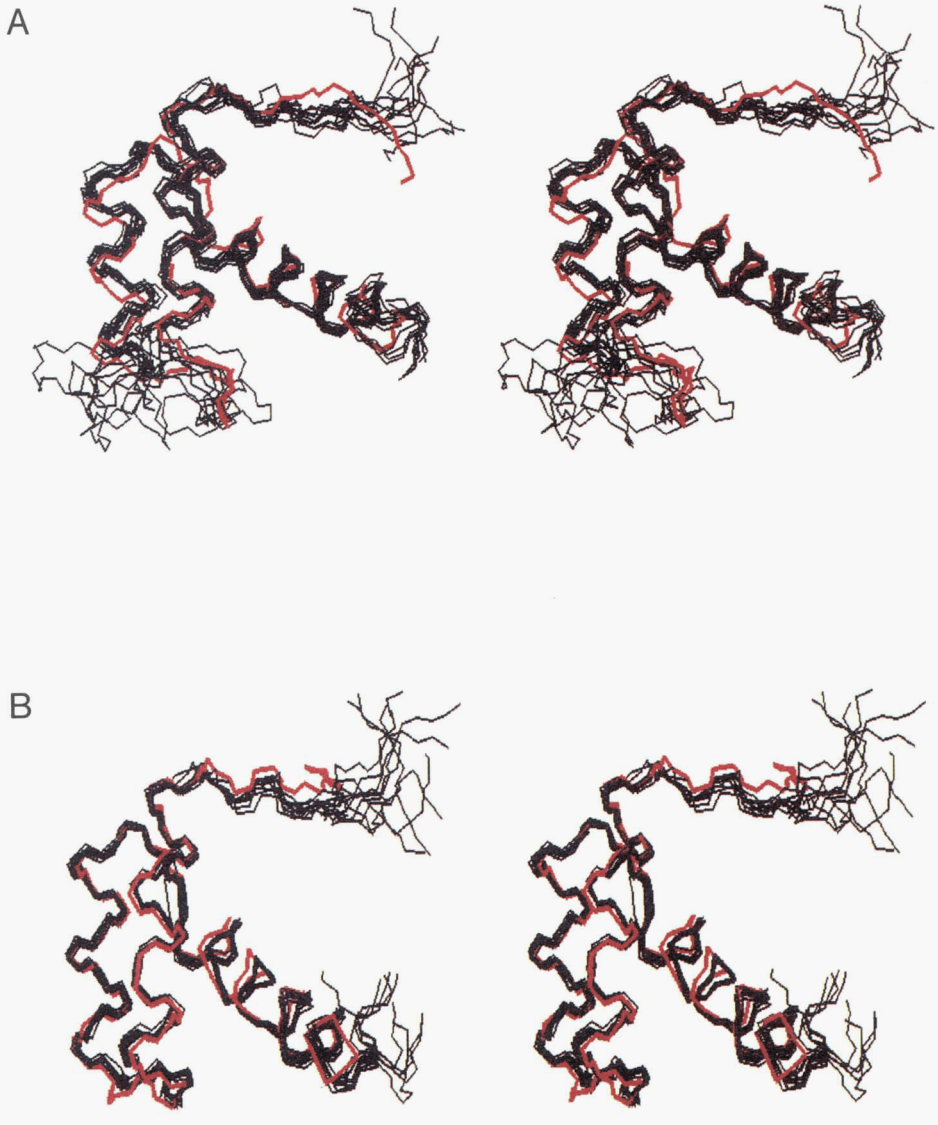

Fig. 2. A: Stereo diagram showing superpositions for backbone atoms of 10 homology-modeled structures of Matα2 together with the X-ray structure (red) from the Matα2-DNA complex (Wolberger et al., 1991). **B:** Stereo diagram showing superpositions for backbone atoms of 10 homology-modeled structures of Antp together with the average NMR structure (red) from the Antp–DNA complex (Billeter et al., 1993).

algorithms (Bruccoleri & Karplus 1987, 1990; Bruccoleri et al., 1988; Bruccoleri & Novotny, 1992) of the CONGEN computer program. By generating a family of structures, one obtains an estimate of the precision of the homology model given the assumptions made regarding homology constraints and the validity of the potential energy function. The accuracy of our methodology was determined by comparisons with experimental crystallographic and NMR structures.

For the homeodomains studied here, a small number of carefully selected α-helical constraints (e.g., enforcing all helical hydrogen bonds) together with several properly chosen tertiary packing constraints would be enough to define the chain fold to ~1 Å. In our approach, the bias of the modeler in selecting such constraints is removed and replaced with a random selection of many more homology constraints (in this case, ~16 constraints per residue) than the minimum number required to mimic the chain fold. In addition, in this work, the distance constraints are defined loosely in order to provide guidance to the information inherent in the CHARMM potential energy function, while allowing the molecule to adopt an energetically relaxed conformation. A more simplified approach excluding conformational energy would not provide one of the key features of our modeling method, i.e., low-energy side-chain packing that results from combination of homology constraints with the energy potential. Moreover, the automated methodology outlined here is not restricted to α-helical proteins and has also been used successfully to homology model β-barrel structures of *Escherichia coli* cold shock proteins (W. Feng, R. Tejero, & G.T. Montelione, unpubl. results).

We have also compared backbone (N, C^{α}, C′) RMS atomic deviations between the experimental structures of Matα2 and Antp and (1) the averaged coordinates of the corresponding "predicted" structures and (2) the engrailed template structure from which the

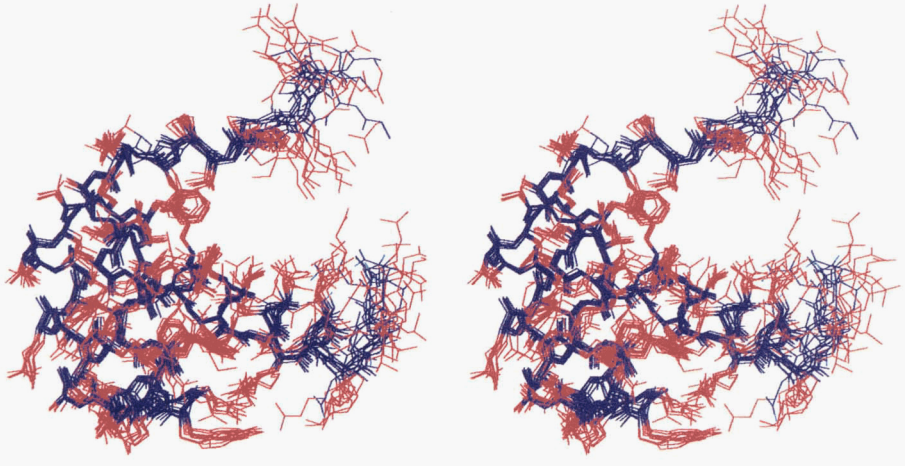

Fig. 3. Stereo diagram showing superpositions of heavy atoms (backbone and side chain) for 10 homology-modeled structures of murine Msx-1.

homology constraints are derived (Table 1). For Matα2, the backbone RMSDs of the X-ray crystal structure to the averaged coordinates of the predicted structure (0.86 Å) and to the engrailed template structure (0.85 Å) are essentially identical. Similarly, for the Antp homeodomain, the backbone RMSDs of the averaged NMR coordinates to the averaged coordinates of the predicted structure (0.76 Å) is only slightly lower than the corresponding RMSD to the engrailed template structure (0.79 Å). In both cases, the backbone conformations of the "predicted" structures are not significantly closer to the actual experimental structures than the backbone conformation of the engrailed template from which they are derived. This is because the extensive network of homology constraints used in this modeling ensures that the homology model will be very similar to the template from which it is derived.

On the other hand, the template structure is not an accurate model of the target structures because is has different side-chain residues. Although potentially providing a reasonable prediction of the backbone conformation, simple mutation of the engrailed structure into the structures of Matα2 or Antp (even if followed by energy minimization), would result in structures that are physically unreasonable and much higher in energy than the structures generated by our procedure. By using simulated annealing methods, "predicted" structures are generated that satisfy *both* the homology constraints and conformational energy constraints imposed by the CHARMM potential energy function. In addition, the SARMD procedure generates a family of structures that provides a measure of the uncertainty of the structure prediction, given the assumption that "homologous atoms" will have similar relative positions in homologous structures. Although using the template structure as the predicted structure provides proper orientation of the homeodomain helices, the challenge addressed by the simulated annealing methodology described here is to generate the *solution space* of low-energy conformations that are consistent with the homology constraints.

Our approach generates the family of structures consistent with a set of "homology constraints" and energy considerations, including steric packing effects. In the crystal structure of the engrailed homeodomain (Kissinger et al., 1990), the N-terminal polypeptide segment interacts with the minor groove of duplex DNA. These steric constraints are not included in our homology modeling calculations. For this reason, the N-terminal polypeptide segments of Matα2, Antp, and Msx1 are not packed against other portions of the protein structure, and are not as tightly defined by the combined homology and energetic constraints as other regions of these homology models (Fig. 2A,B).

In the case of insertions and deletions, we also used the conformational search algorithms of CONGEN (Bruccoleri & Karplus, 1987; Bruccoleri, 1993) to construct low-energy polypeptide loop conformations, as described in the Materials and methods. In Matα2, we defined residues 23, 23a, 23b, 23c, 24, 25, 26, and 27 as an unconstrained loop. Conformational searching (Bruccoleri, 1993) allowed us to explore the entire conformational space and find conformations with the lowest energies for this octapeptide loop. The search resulted in two families of loop backbone structures (Fig. 2A), one of which includes the X-ray structure (Wolberger et al., 1991). This result indicates that there might in fact be more than one low-energy conformation for this surface loop, although the electron density apparently could be fit to a single backbone conformation.

Comparison with other approaches using satisfaction of homology constraints

This section compares our hybrid SARMD search method using CONGEN with related automated methods for spatial satisfaction of homology constraints. The comparison focuses on the methods used for generating structures and the kinds of homology and conformational energy information that are used. No efforts have been made to date to compare relative performance or reliability of these various approaches.

Among the several general methods available for homology modeling, approaches that are most similar to ours include those using DG calculations with homology constraints (Havel & Snow, 1991; Havel, 1993; Srinivasan et al., 1993; Sudarsanam et al., 1994), automated methods using restrained simulated annealing with selected spatial constraints judged to be important for the fold and/or function (Brocklehurst & Perham, 1993), and the probability density function approach employed by the program MODELLER (Sali & Blundell, 1993). Like our approach, the basic philosophy of these methods is to generate homology models au-

tomatically with little or no user intervention. Significant differences among these various approaches and our method include (1) details of algorithms used for structure generation, (2) the target functions and their relationships to homology structural information, and (3) the use of a set of superimposed conformers to interpret the consistency of the homology modeling prediction in different parts of the protein structure.

The DG approaches (Havel & Snow, 1991; Havel, 1993; Srinivasan et al., 1993; Sudarsanam et al., 1994) use metric-matrix methods to embed the structure into three-dimensional space, while minimizing homology and steric constraint violations. The method of Sali and Blundell (1993) uses a variable-target function optimization procedure with conjugate gradient minimization followed by simulated annealing with molecular dynamics. It is unique in that homology-constraint information is evaluated in the target function in the form of a probability density function that describes various homologous features to be restrained. This continuous probability density function differs from the flat-bottomed upper/lower bound constraints used by DG and restrained MD methods. Previously described pure SARMD approaches have used a knowlege-based selection of a relatively small number of homology constraints (Brocklehurst & Perham, 1993) or have used only homology constraints between backbone atoms and minimal conformational energy information (Engh et al., 1990; Fujiyoshi-Yoneda et al., 1991). The procedures of Havel (1993) and Sali and Blundell (1993) also use molecular dynamics to refine structures that are first generated by DG or variable-target function gradient minimization methods. Although the energies of homology models produced by these methods are improved by the restrained MD calculations (Havel, 1993; Sali & Blundell, 1993), these conformers may be trapped in local energy minima that cannot be overcome without a sufficiently long high-temperature simulated annealing procedure. As a result, the family of structural models may retain energetically unfavorable features, and may not sample adequately the range of conformations that are consistent with the assumptions of the homology modeling.

In our approach, a subset of homology constraints are selected randomly without user intervention from the "homologous distances" in the template structure. SARMD with CONGEN is then used right from the start to generate a family of structures that satisfies both homology and energetic constraints while maximizing the sampling of the solution space. The convergence rate of these SARMD calculations is enhanced by using the flexible distance restraint function of CONGEN, as described by Bassolino-Klimas et al. (1996). Efficient sampling of the solution space of side-chain and loop conformations is provided both by the simulated annealing protocol and by the subsequent directed conformational search. Like the procedures of Brocklehurst and Perham (1993), Havel (1993), and Sali and Blundell (1993), our method generates families of homology-modeled structures that are used to evaluate the consistency of the structure prediction in different parts of the model.

Further improvements in the CONGEN SARMD/search approach

Each of the methods described above have their respective strengths and weaknesses, and it may be advantageous to incorporate ideas described by other workers into our CONGEN homology modeling calculations. The work described by Havel (1993) used multiple template structures for generating homology constraints. The pdf's used in the method of Sali and Blundell (1993) are also derived from multiple homologous structures. In this study, we used a single template structure (i.e., the crystal structure of the engrailed protein), although it is also possible to define homology constraints for SARMD from multiple template structures in the future. In addition, other approaches (Brocklehurst & Perham, 1993; Havel, 1993; Sali & Blundell, 1993) have included dihedral angle homology constraints, hydrogen bond homology constraints, and/or explicit water molecules in their homology modeling calculations, all of which might be beneficial to include in our SARMD/search approach. Future homology modeling applications of SARMD/search with CONGEN could also benefit from more realistic descriptions of electrostatics, solvent effects, counter-ion effects, and other interactions between the modeled protein and other molecules. In the case of these homedomains, the template engrailed structure comes from the structure of a complex between engrailed and duplex DNA (Kissinger et al., 1990); modeling of the entire protein–DNA complex would provide better definition and more correct predictions of conformations of interfacial side chains because the interactions with DNA would act to reduce the number of isoenergetic side-chain conformations.

Despite these possible shortcomings, the SARMD/search method using CONGEN with homology constraints provides a robust and reliable approach for homology modeling of small single-chain proteins. The predicted homeodomain structures described here exhibit precision and accuracy similar to that available from medium-resolution solution NMR structures. We have also had good success using this approach for homology modeling β-barrel protein structures (W. Feng, R. Tejero, & G.T. Montelione, unpubl. results). Overall, the results presented in this paper provide a good example of the utility of homology modeling using spatial homology constraints (Havel & Snow, 1991) applied to an important class of transcription-regulating domains. The family of modeled structures of the Msx-1 homeodomain has been deposited in the Brookhaven Protein Data Bank for use in future comparisons of the predicted structure with NMR and X-ray structures once they are available or in molecular replacement approaches for determining its structure by these experimental methods.

Predicted structure of murine Msx-1

Figure 4 shows a ribbon model depicting the average, energy-minimized predicted structure of Msx-1. The structure consists of three α-helices: helix I (residues 10–22), helix II (residues 28–37), and helix III (residues 42–58). The helix content in the predicted structure is ~67% and is consistent with estimates from analysis of far-UV CD spectra of ~70% helix content in aqueous solution at pH 7.0 and temperature of 22 °C (Shang et al., 1994). Helices I and II are antiparallel, whereas helix III is perpendicular to the other two helices. The N-terminal polypeptide segment (residues 1–9) is in an extended conformation directed away from the three helical bundle. The conformation of the segment, modeled from the bound-state conformation of the engrailed–DNA complex, is not defined uniquely in our simulated annealing calculations (Fig. 3) and may, in fact, be quite flexible in Msx-1 when it is not bound to DNA.

Dihedral angle order parameters $S(\phi)$ and $S(\psi)$ (Hyberts et al., 1992) and Ramachandran plots computed for this family of homology-modeled Msx-1 structures are shown in Figures 5 and 6, respectively. Only residues of well-defined regions of the structure, defined as those with $S(\phi) + S(\psi) > 1.5$, are shown on the Ramachandran plots. All of these well-defined residues in all of the

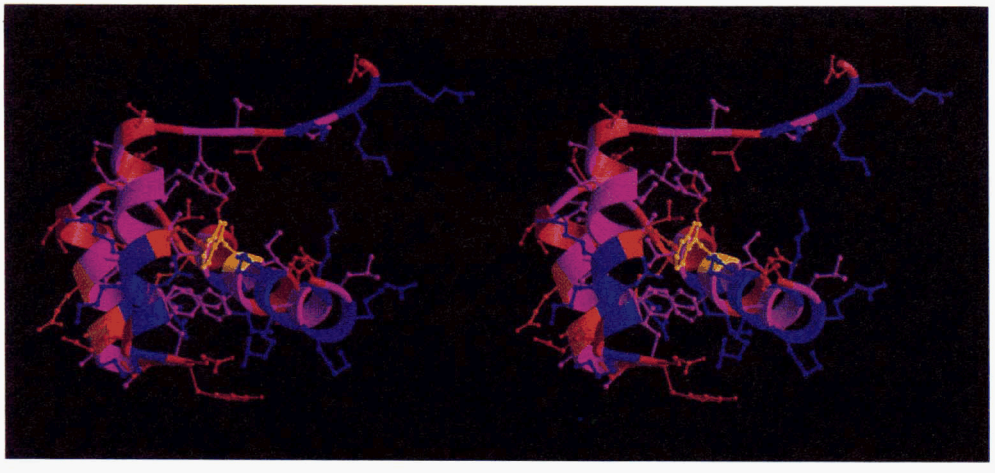

Fig. 4. Ribbons (Carson, 1991) model of the averaged, energy-minimized structure of Msx-1 predicted by homology modeling. The residues are shown with as follows: basic residues, blue; acidic residues, orange; hydrophobic residues, magenta; hydrophilic residues, red; tryptophan, yellow.

models are in low-energy regions of the Ramachandran plot (Fig. 6). The only residue on the right side (i.e., $\phi > 0$) of the Ramachandran map is residue Ser 39, with $\phi_{39} \approx 50°$ and $\psi_{39} \approx 60°$. Because residue Leu 38 has ϕ, ψ values around $-90°$ and 0°, respectively, the polypeptide segment Ser 37–Leu 38–Ser 39–Leu 40 forms a type IV β bend (Lewis et al., 1973) in the predicted structure of Msx-1. Position 39 is occupied by a glycine residue in the template-engrailed protein structure (Kissinger et al., 1990). It will be interesting to learn if this prediction of a positive ϕ value for residue Ser 39 is borne out by future NMR and/or crystal structures of Msx-1 or if it is an artifact of our homology modeling procedure.

As illustrated in the stereodiagram of Figure 3, most of the conformations of side chains in the interior of Msx-1 are tightly restrained by the combination of distance and energy constraints, whereas surface side-chain conformations are more poorly defined in the homology model, as would be expected from the energy potential. Similar distributions of side-chain conformations were observed in the ensembles of homology-modeled Matα2 and Antp structures. This kind of graphical information on the precision of side-chain conformations in these homology models is available only because the models are represented as ensembles of structures and because MD and search methods are used to sample extensively the solution space consistent with the homology and energy constraints.

Conserved residues, located mainly at the helical interfaces of the three-helical bundle fold, form the hydrophobic core of the Msx-1 homeodomain. This core includes many highly conserved hydrophobic residues (Scott et al., 1989) with an invariant Trp 48 at the center (Fig. 4). Near-UV CD spectroscopy and tryptophan fluorescence quenching experiments on Msx-1 demonstrate that residue Trp 48 is, in fact, buried in the hydrophobic core of the native structure of Msx-1 (Shang et al., 1994), as it is in our predicted homology models.

In the engrailed, Antp, and the Matα2 homeodomain–DNA complexes, residues in helix III contact the major groove, residues in the N-terminal arm contact the minor grove, and residues in helix II contact the phosphate backbone (Kissinger et al., 1990; Wolberger et al., 1991; Billeter et al., 1993; Li et al., 1995). It is likely that these contacts will also be found in the Msx-1–DNA complex. The positively charged side chains of residues arginine and lysine (shown in blue in Fig. 4) on the N-terminal arm and on the surface of helix III of Msx-1 are available to interact with DNA in the minor and major grooves, respectively. The highly charged N-terminal arm and third helix can provide high-affinity electrostatic interactions with the polyphosphate backbone of DNA and the complex hydrophobic and ionic nature of this surface can provide specificity to dock the protein at its correct binding position.

A conserved network of salt bridges in the homeodomain chain fold

A network of ionic interactions preserves the interhelical packing of the homeodomain chain fold in Msx-1. In analyzing our predicted structure of Msx-1, we observed a network of several surface salt bridges (Table 3) that appears to be conserved in other homeodomain structures. These are illustrated on the three-dimensional structure of Msx-1 in Figure 7. This network of surface salt bridges in the homology-modeled structure of Msx-1 results primarily from constraints imposed by its predicted chain fold and in part as a

Table 3. *Postulated salt bridges in homology-modeled structures of the Msx-1 homeodomain*

Anion	Cation	Frequency of occurrence[a]
Glu 30 $C^{\delta}OO^{-}$	Lys 19 $N^{\zeta}H_3^{+}$	1.0
Glu 30 $C^{\delta}OO^{-}$	Lys 23 $N^{\zeta}H_3^{+}$	1.0
Glu 33 $C^{\delta}OO^{-}$	Lys 19 $N^{\zeta}H_3^{+}$	0.7
Glu 42 $C^{\delta}OO^{-}$	Arg 31 $N^{\eta}H_2^{+}$	0.9
Glu 17 $C^{\delta}OO^{-}$	Arg 18 $N^{\eta}H_2^{+}$	0.9
Glu 17 $C^{\delta}OO^{-}$	Arg 52 $N^{\eta}H_2^{+}$	0.7
Glu 17 $C^{\delta}OO^{-}$	Arg 21 $N^{\eta}H_2^{+}$	0.6

[a]Criteria used to identify postulated salt bridges are outlined in Materials and methods. The frequency of occurrence is the fraction of the 10 homology-modeled structures containing the postulated salt bridge. Potential salt bridges with frequency of occurrence <0.6 are not reported.

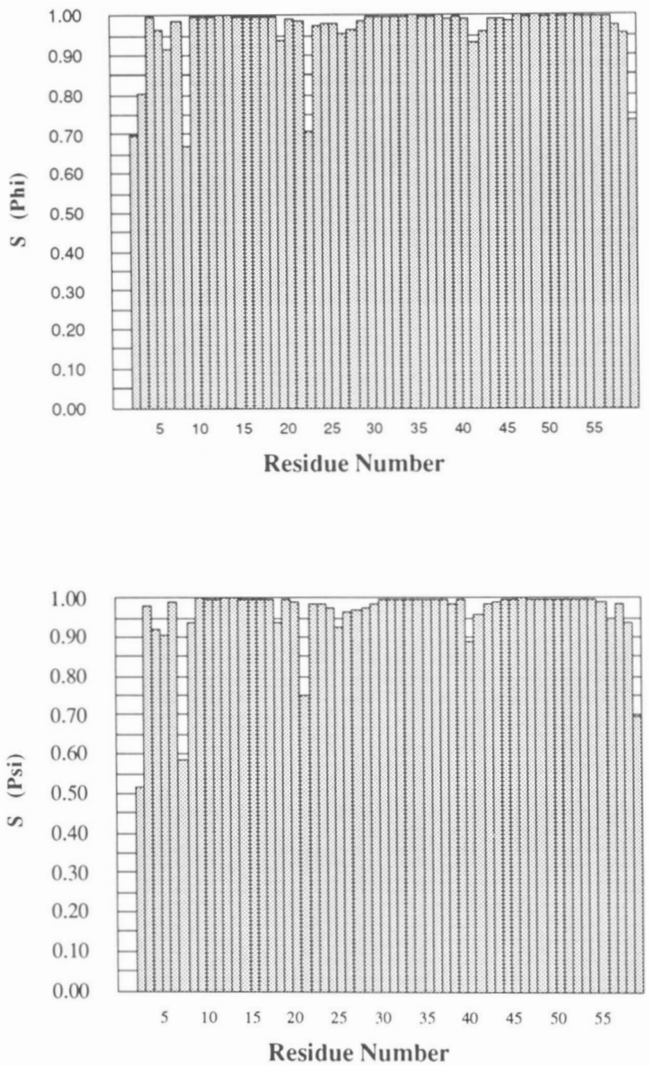

Fig. 5. Backbone dihedral angle order parameters (Hyberts et al., 1992) for the ensemble of homology-modeled structures of Msx-1.

consequence of the CHARMM force field, which drives oppositely charged atoms toward one another.

In the Msx-1 homeodomain structure, the ionizable side-chain atoms of residues Lys 19, Lys 23, and Glu 30 are spatially very close together, forming salt bridges between helix I and helix II in all the 10 modeled structures. Residues Lys 19 and Lys 23 are four residues apart in the C-terminus of helix-I and their side chains are on the same side of the helix and adjacent to one another in space. Salt bridges between positions 19 and/or 23 in helix I and position 30 in helix II are also present in other homeodomains, including the Glu 19–Arg 30 salt bridge in both the engrailed and Antp homeodomains, the Glu 22–Arg 30 salt bridge in engrailed, and the "inverse" Arg 19–Glu 30 salt bridge in the predicted structure of thyroid transcription factor 1 (TTF-1) (Fogolari et al., 1993). Although a homologous salt bridge is not present in Matα2, in the structure of the related yeast, Mat-a1 homeodomain includes a salt bridge between residues Lys 23 and Glu 30 (Li et al., 1995). The double mutant [Lys 19-Glu, Glu 30-Arg] Msx-1, with an "inverse" salt bridge, exhibits a native-like DNA-binding affinity (Isaac et al., 1995).

The side chains of residues Arg 31 and Glu 42 also form a salt bridge between helix II and helix III in 9 of 10 modeled Msx-1 structures (Fig. 7; Table 3). Homologous 31–42 salt bridges are also present in the engrailed and Antp structures. Although Matα2 does not have a salt bridge between 31 and 42, it does have the "inverse" Glu 32–Arg 42 salt bridge. These compensating amino acid substitutions in Matα2 and the strong conservation of the 31/32–42 ionic interaction suggest that this salt bridge between helix II and helix III has an important structural and/or functional role.

Another frequently observed salt bridge is Glu 17–Arg 52, which connects helix I and helix III; it is found in 7 of 10 homology-modeled structures (Table 3) and is conserved in the Antp and Matα2 homeodomain structures. In the TTF-1 homeodomain, a salt bridge between Glu 17 and His 52 is also predicted (Fogolari et al., 1993). However, a homologous 17–52 salt bridge is not present in engrailed homeodomain, where both residues are lysines.

In a covariance analysis of 60 homeodomain protein sequences, Clarke (1995) has observed that residue pairs 19/30, 31/42, and 17/52 are among the most strongly correlated covariant residue pairs. Moreover, the nature of these covariances generally functions to preserve salt bridges at these three surface sites (Clarke, 1995). The energetic basis of the requirement for these surface salt bridges is not yet certain. Residues Arg 31 of engrailed and Arg 42 of Matα2 each make contacts with corresponding phosphate atoms in these homeodomain–DNA complexes, suggesting that there may be a functional constraint requiring a basic residue at one (but not both) of these two sites (Clarke, 1995). However, residue pairs 19/30 and 17/52 are distant from the DNA-binding site and do not interact with DNA in the structures of complexes.

Considering that solvent-accessible surface ionic interactions will be suppressed by the high dielectric of bulk water and by counterion salts, it is unlikely that these salt bridges contribute significantly to the thermodynamic stability of the free homeodomain. This suggests that the strong conservation of these salt bridges reflects a requirement to balance charges because these faces of the homeodomain (Fig. 7) are buried in protein–protein and protein–nucleic acid interactions that occur in the formation of functional transcription complexes. For example, in the crystal structure of the Mata1/Matα2 homeodomain heterodimer bound to DNA, the Lys 23–Glu 30 salt bridge of Mata1 is buried at the interface between Mata1 and the carboxy-terminal tail of Matα2 (Li et al., 1995). Because it is energetically very unfavorable to bury an unbalanced charge in a protein–protein interface, the presence or absence of these generally conserved salt bridges on the surfaces of different homeodomains potentially can play an important role in modulating the energetics of interactions among homeodomains and between homeodomains and other proteins and/or DNA in transcription complexes.

Implications of the Msx-1 structure for protein engineering

Using structural information derived from this homology model of Msx-1, we have designed and synthesized homeodomain Ala-Msx (Shang et al., 1994), which contains most homeodomain consensus residues (Scott et al., 1989), but replaces most nonconsensus residues with alanine. Ala-Msx contains 46% Ala and binds to duplex DNA with a sequence-specific dissociation constant $K_D < 100$ nM (Shang et al., 1994). This alanine-substituted analogue of Msx-1, which retains the hydrophobic core defined by the homology model

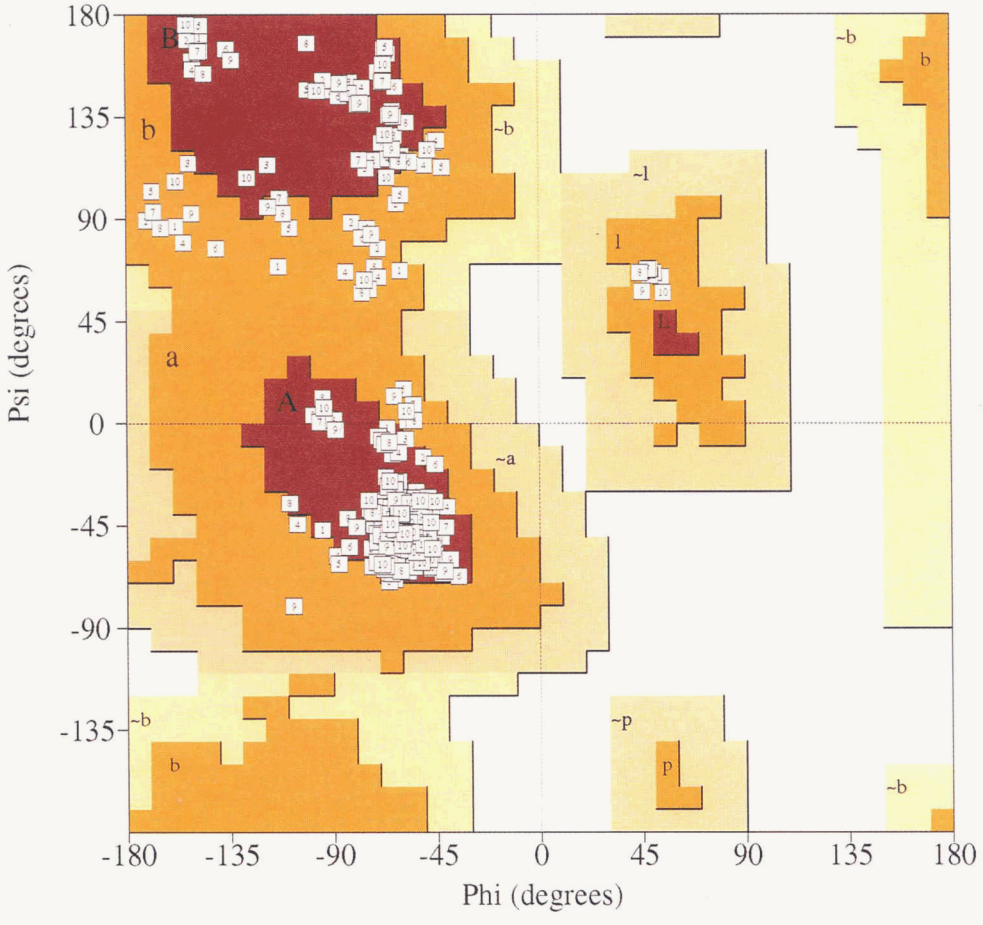

Fig. 6. Composite PROCHECK-NMR Ramachandran plot for the family of homology-modeled structures of Msx-1. Only residues with backbone dihedral angle order parameters $S(\phi) + S(\psi) < 1.5$ are plotted (i.e., excluding residues 1, 2, 57, 58, and 59; see Fig. 5).

of Figure 3 and the key salt-bridges shown in Figure 7, has ~70% α-helical content based on far-UV CD spectroscopy. However, NMR and fluorescence studies indicate that Ala-Msx is a relatively dynamic protein with a transient tertiary structure (Shang et al., 1994). These studies demonstrate the value of modeling calculations in the initial stages of a protein structure engineering project, where only the overall chain fold and some key details of tertiary structure are required.

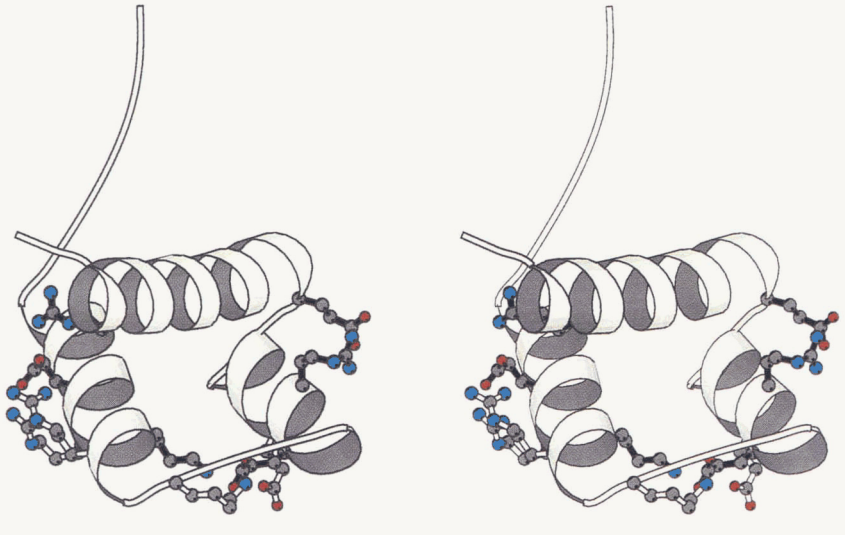

Fig. 7. Network of conserved surface salt bridges identified in the homology-modeled structure of Msx-1.

Materials and methods

Homology modeling algorithm

We have developed an automated hybrid method of homology modeling using simulated annealing with restrained molecular dynamics calculations and conformational search (SARMD/search) with the molecular mechanics program CONGEN (Bruccoleri & Karplus, 1987, 1990; Bruccoleri, 1993; Bassolino-Klimas et al., 1996). Our procedure generates a family of three-dimensional protein structures for a homologous protein based on the atomic coordinates of a known protein structure. The one (or more) known structures upon which the modeling is based is referred to as the "template structure." "Homologous atoms" are defined using the alignments of amino acid sequences of the template protein and the protein structure to be predicted; only heavy atoms (atoms other than hydrogen) are considered in defining "homologous atoms" (Fig. 8). If, at a certain position, the amino acid residues are conserved in both known and unknown structures, all the heavy atoms of that residue are "homologous atoms"; if the residues differ, then only the atoms that are the same chemical type and hybridization state are defined as "homologous atoms." For example, Cβ atoms in Ala and Ser are homologous, because both are at the β position and have sp^3 hybridized atomic orbitals. However, none of Cγ atoms of long side chains (e.g., Lys, Val, Leu, and Ile) are homologous to Cγ atoms of the aromatic amino acids (i.e., Phe, Tyr, His, and Trp) because the Cγ atomic orbitals in these two classes are hybridized differently. Methyl, methylene, and methine carbons that have the same position in the side chain are considered to be homologous because they have the same sp^3 hybridization, even if they are attached to heteroatoms. Homology constraints involving nonidentical but homologous atom substitutions at chiral Cβ sites of Ile and Thr for pro-chiral isopropyl methyl groups of Leu and Val (or visa versa) require special considerations. In these calculations, these special situations did not occur.

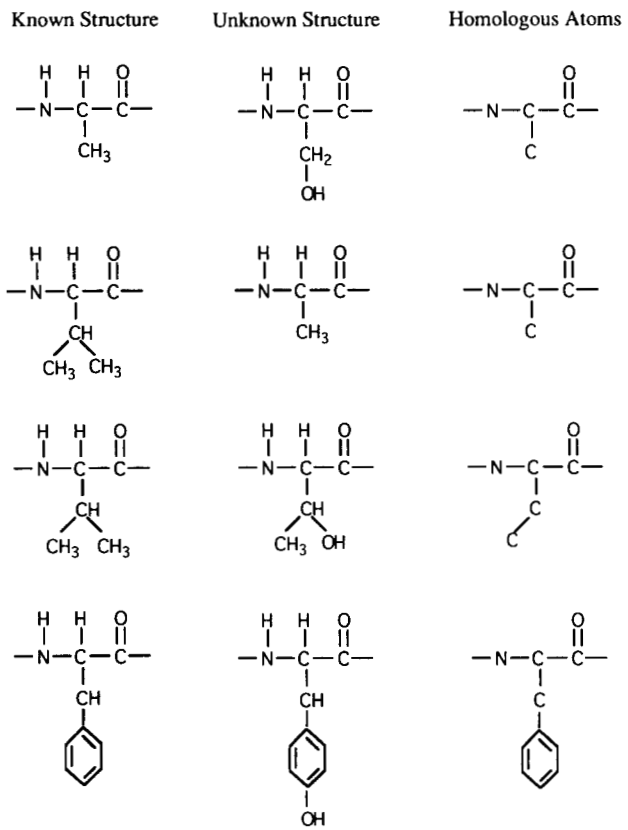

Fig. 8. Representative examples of approach used to define homologous atoms between aligned residues.

The definitions of homologous atoms and the success of the method rely on a proper alignment of homologous sequences. For the homeodomains studied here, this was quite straightforward (Fig. 1). In cases of insertions or deletions of polypeptide segments, no "homologous atoms" are defined in the inserted or deleted region. Moreover, because the local conformation is likely to change in order to accommodate insertion or deletion of the polypeptide chain, we also do not define homologous atoms in the vicinity of (about three residues before and after) the insertion or deletion.

Next, the atomic coordinates of the "template" structure are used to compute distances between these "homologous atoms." For a pair of homologous 60-residue proteins, there are about 400 "homologous atoms," and some 80,000 "homologous distances." From these, about 1,000 were selected randomly and used to create pairs of upper- and lower-bound "homology constraints." "Homology constraints" which do not restrict intervening dihedral angles were excluded. Interaction maps derived from the corresponding constraint files, summarizing the distributions of these constraints in the sequences of Matα2, Antp, and Msx-1, are presented in the Electronic Appendix. Visual examination of these interaction maps demonstrates the random nature of the constraint selection. From each "homologous distance constraint," upper and lower bounds are created by adding and subtracting, respectively, 10% of the exact distance. These upper- and lower-bound homology constraints are then used in structure generation calculations. SARMD (Bassolino-Klimas et al., 1996; Tejero et al., 1996) is then used to generate a family of structures that satisfies the homology constraints. This density of constraints (~16.7 constraints per residue) was chosen to be comparable to the number of NOE distance constraints obtained normally in high-resolution NMR structure determinations. Although, to date, no efforts have been made to correlate homology constraint density with the accuracy and precision of the homology model, we have shown previously that constraint densities of 16 per residue are sufficient for generating accurate and precise structures of the crambin protein in tests executed to validate the restrained minimization functions of CONGEN (Bassolino-Klimas et al., 1996).

Simulated Annealing with Restrained Molecular Dynamics

Three-dimensional structures are computed from the homology distance constraints using the CONGEN program (Bruccoleri & Karplus, 1990; Bassolino-Klimas et al., 1996). A target function describing the conformational energy and residual "homology constraint" violations is minimized by simulated annealing with molecular dynamics. Conformational energies are computed using the CHARMM force field (Brooks et al., 1983) and distance-dependent dielectric constant ($\epsilon = r$). In our current implementation, starting structures are fully extended conformations of the unknown protein differing in random assignments of atomic velocities.

Prior to the simulated annealing calculations, an unrestrained adopted-basis Newton-Raphson (ABNR) minimization is performed on the starting conformations in order to relieve any en-

ergetically unfavorable contacts. The annealing procedure includes two stages: weight annealing and temperature annealing (Bassolino-Klimas et al., 1996; Tejero et al., 1996). In weight annealing, restrained MD calculations are obtained at 1,000 K while gradually increasing the relative weight on the homology constraint term in the target function, K_{homol}, from 0 to 100 kcal/mol-Å^2. This stage includes 21 MD periods of variable lengths (10 MD periods of 4 ps each plus 11 periods of 3 ps each). Next, temperature annealing is conducted by computing the restrained MD trajectory (with K_{homol} = 100 kcal/mol-Å^2) while slowly cooling the system from 1,000 K to 300 K in decrements ranging from 100 to 5 K. This process includes 12 MD periods of 1 ps each, corresponding to a total temperature annealing time of 12 ps. Next, the system is equilibrated by continuing the restrained MD trajectory for 5 ps with temperature rescaling at 300 K. Finally, a 10-ps restrained MD trajectory (without temperature rescaling) is carried out. The average structure sampled during the last 2 ps of this final MD trajectory is computed and then restrained-energy minimized using the ABNR method. In these calculations, all peptide bonds were kept fixed in the planar *trans* conformation using a weighted flat-bottomed hyperbolic restraint function (K_{dih} = 200 kcal/mol-deg) with a minimum at 180 ± 3°. Each structure generated by simulated annealing required approximately 6 h of cpu time on an R4000 processor of a SGI workstation. The software for automatic generation of CONGEN homology constraint files from a pair of aligned protein sequences, the SARMD protocol files used in this work, and the CONGEN program itself are available from the authors.

Selection of 10 "best" conformers

For each homeodomain, 25 conformers were computed by SARMD with CONGEN. The 20 that best satisfied the homology constraints were then selected. Of these, the 15 with lowest values of Van der Waals energy were then identified, and from these, the 10 with lowest values of conformational energy (including electrostatics) were selected to represent the predicted structure of the homeodomain. These 10 "best" structures usually corresponded to the 10 that best satisfied the homology constraints.

Conformational search

The construction of the models of Antp and Msx-1 also used the side-chain conformational search capabilities of CONGEN. The model for Matα2 utilized both side-chain and loop conformational search methods. Side-chain modeling was performed using the iterative side-chain method in CONGEN (Bruccoleri & Karplus 1987, 1990). In this method, the conformation of each side chain is minimized using exhaustive conformational search while the other side chains are fixed. The search is performed repetitively over all side-chain χ angles that involve nonhomologous atoms until the energy converges. The chi angles are sampled using a grid of 30°, except for lysines and arginines containing no homologous side-chain atoms, where a 60° grid was used. This grid was used because of the large number of possible side-chain conformers in these long side chains. Van der Waals avoidance was used with a cutoff of 5 kcal/mol. Side-chain constructions required a few minutes of CPU time per protein on a single R4000 processor of an SGI workstation.

In the case of Matα2, the interhelical loop between residues Asn 23 and Asp 27 was constructed for each model using protocols similar to those described previously (Bruccoleri et al., 1988; Bruccoleri & Novotny, 1992). Dihedral angles in the octapeptide segment Asn 23–Ile 23a–Glu 23b–Asn 23c–Pro 24–Tyr 25–Leu 26–Asp 27 were sampled using a 30° grid and a van der Waals cutoff of 20 kcal/mol. Only backbone conformations within 2 kcal/mol of the minimum on the Ramachandran map were used (Bruccoleri & Karplus, 1987). The loop was closed using the modified Gō and Scheraga algorithm (Gō & Scheraga, 1970; Bruccoleri & Karplus, 1985). The conformations of loop side chains were constructed by the search method using van der Waals avoidance with a 5 kcal/mol cutoff and minimum energy periodicity grid (120° for torsions over sp^3 hybridized atoms and 180° for torsions involving sp^2 hybridized atoms).

For eight of the models, this procedure generated many low-energy conformations; however, CONGEN was unable to find low-energy conformations for the loop in two models. In one case, the Ramachandran map cutoff was raised to 5 kcal/mol. In the other case, the carboxylate atoms of residue Glu 32 were also included in the search because of close contacts to other atoms in the loop. In addition, the Ramachandran map cutoff was raised to 5 kcal/mol. The loop constructions required elapsed times of between 5 min and 21 h on seven R4000 processors of an eight processor SGI workstation.

Salt bridge identification

For defining salt bridges, the electrostatic energy was assumed to be greater than the translational and rotational thermal energy, E_T = 5/2 RT, which corresponds to 1.5 kcal/mol. For calculating the electrostatic energy, we assume that each oxygen atom of a carboxylate group or terminal nitogen atom of a guanido group contributes charge $\pm e/2$, whereas ammonium nitrogens have charge $+e$, and use the distance between changes in Ångstroms as the electrostatic constant, ϵ. From these criteria, we determined that a significant salt bridge interaction occurs between two oppositely charged ionic side chains when the distance between the two closest oppositely charged atoms is less than 7.5 Å. Accordingly, we used 7.0 Å as a cutoff distance for defining a salt bridge between oppositely charged atoms in our structural analysis. Although this analysis is more sophisticated than is justified, particularly considering the lack of explicit solvent and counter ion effects in these calculations, a 7.0-Å cutoff is a reasonable definition for these surface ionic interactions.

Protein structure coordinates

PDB coordinate files for Antennapedia (1AHD) and engrailed (1HDD) homeodomains were taken from the Brookhaven Protein Data Bank. Coordinates for the Matα2 homeodomain were kindly provided by Prof. Cynthia Wolberger.

Supplementary material in Electronic Appendix

Three figures in Adobe Postscript format showing interaction maps of the distributions of intra- and interresidue homology constraints used for homology modeling the structures of the Matα2, Antp, and Msx-1 homeodomains are found in the Electronic Appendix.

Acknowledgments

We thank R. Watson for useful editorial comments on the manuscript. This work was supported by grants from the National Institutes of Health (GM-47014, GM-50733), the National Science Foundation (MCB-9407569), a

National Science Foundation Young Investigator Award (MCB-9357526), and a Camille Dreyfus Teacher-Scholar Award. Support for computing facilities was provided by a grant from the W.M. Keck Foundation. R.T. also acknowledges partial support from the University of Valencia.

References

Assa-Munt N, Mortishire-Smith RJ, Aurora R, Herr W, Wright PE. 1993. The solution structure of the Oct-1 POU-specific domain reveals a striking similarity to the bacteriophage λ repressor DNA-binding domain. *Cell 73*:193–205.

Bajorath J, Stenkamp R, Aruffo A. 1993. Knowledge-based model building of proteins: Concepts and examples. *Protein Sci 2*:1798–1810.

Bassolino-Klimas D, Tejero R, Krsystek SR, Metzler WJ, Montelione GT, Bruccoleri RE. 1996. Simulated annealing with restrained molecular dynamics using a flexible restraint potential: Theory and evaluation with simulated NMR constraints. *Protein Sci 5*:593–603.

Bazan JF. 1990. Structural design and molecular evolution of a cytokine receptor superfamily. *Proc Natl Acad Sci USA 87*:6934–6938.

Billeter M, Güntert P, Luginbühl P, Wüthrich K. 1996. Hydration and DNA recognition by homeodomains. *Cell 85*:1057–1065.

Billeter M, Qian YQ, Otting G, Müller M, Gehring W, Wüthrich K. 1993. Determination of the nuclear magnetic resonance solution structure of an *Antennapedia* homeodomain–DNA complex. *J Mol Biol 234*:1084–1093.

Blundell TL, Sibanda BL, Pearl L. 1983. Three-dimensional structure, specificity and catalytic mechanism of renin. *Nature 304*:273–275.

Brocklehurst SM, Perham RN. 1993. Prediction of the three-dimensional structures of the biotinylated domain from yeast pyruvate carboxylase and of the lipoylated H-protein from the pea leaf glycine cleavage system: A new automated method for the prediction of protein tertiary structure. *Protein Sci 2*:626–639.

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. 1983. CHARMM: A program for macromolecular energy minimization and dynamics calculations. *J Comput Chem 4*:187–217.

Bruccoleri RE. 1993. Application of systematic conformational search to protein modeling. *Mol Sim 10*:151–174.

Bruccoleri RE, Haber E, Novotny J. 1988. Structure of antibody hypervariable loops reproduced by a conformational search algorithm. *Nature 335*:564–568.

Bruccoleri RE, Karplus M. 1987. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers 26*:137–168.

Bruccoleri RE, Karplus M. 1990. Conformational sampling using high-temperature molecular dynamics. *Biopolymers 29*:1847–1862.

Bruccoleri RE, Novotny J. 1992. Antibody modeling using the conformational search program CONGEN. *Immunomethods 1*:96–106.

Carson M. 1991. RIBBONS 2.0. *J Appl Crystallogr 24*:958–961.

Catron KM, Iler N, Abate-Shen C. 1993. Nucleotides flanking a conserved TAAT core dictate the DNA-binding specificity of three murine homeodomain proteins. *Mol Cell Biol 13*:2354–2365.

Catron KM, Wang H, Hu G, Shen MM, Abate-Shen C. 1996. Comparison of Msx-1 and Msx-2 suggests a molecular basis for functional redundancy. *Mech Dev 55*:185–199.

Catron KM, Zhang H, Marshall SC, Inostroza JA, Wilson JM, Abate C. 1995. Transcriptional repression by Msx-1 does not require homeodomain DNA-binding sites. *Mol Cell Biol 15*:861–871.

Ceska TA, Lamers M, Monaci P, Nicosia A, Cortese R, Suck D. 1993. The X-ray structure of an atypical homeodomain present in the rat liver transcription factor LFB1/HNF1 and implications for DNA binding. *EMBO J 12*:1805–1810.

Chothia C, Lesk AM, Levitt M, Amit AG, Mariuzza RA, Phillips SEV, Poljak RJ. 1986. The predicted structure of immunoglobulin D1.3 and its comparison with the crystal structure. *Science 233*:755–758.

Claessens M, Cutsem EV, Lasters I, Wodak S. 1989. Modeling the polypeptide backbone with "spare parts" from known protein structures. *Protein Eng 2*:335–345.

Clarke ND. 1995. Covariation of residues in the homeodomain sequence family. *Protein Sci 4*:2269–2278.

Cox M, Dekker N, Boelens R, Verrijzer CP, van der Vliet PC, Kaptein R. 1993. NMR studies of the POU-specific DNA-binding of Oct-1: Sequential ^{1}H and ^{15}N assignments and secondary structure. *Biochemistry 32*:6032–6040.

Davidson DR, Crawley A, Hill RE, Tickle C. 1991. Position-dependent expression of two related homeobox genes in developing vertebrate limbs. *Nature 352*:429–431.

Dessain S, Gross TC, Kuziora MA, McGinnis W. 1992. Antp-type homeodomains have distinct DNA binding specificities that correlate with their different regulatory functions in embryos. *EMBO J 11*:991–1002.

Dubnau J, Struhl G. 1996. RNA recognition and translational regulation by a homeodomain protein. *Nature 379*:694–699.

Ekker SC, von Kessler DP, Beachy PA. 1992. Differential DNA sequence recognition is a determinant of specificity in homeotic gene action. *EMBO J 11*:4059–4072.

Engh RA, Wright HT, Huber R. 1990. Modeling of the intact form of the α-proteinase inhibitor. *Protein Eng 3*:469–477.

Fogolari F, Esposito G, Viglino P, Damante G, Pastore A. 1993. Homology model building of the thyroid transcription factor 1 homeodomain. *Protein Eng 6*:513–519.

Fujiyoshi-Yoneda T, Yoneda S, Kitamura K, Amisaki T, Ikeda K, Inoue M, Ishida T. 1991. Adaptability of restrained molecular dynamics for tertiary structure prediction: Application to *Crotalus atrox* venom phospholipase A_2. *Protein Eng 4*:443–450.

Gehring WJ. 1987. Homeo boxes in the study of development. *Science 236*:1245–1252.

Gehring WJ, Qian YQ, Billeter M, Furukubo-Tokunaga K, Schier AF, Resendez-Perez D, Affolter M, Otting G, Wüthrich K. 1994. Homeodomain DNA recognition. *Cell 78*:211–223.

Gō N, Scheraga HA. 1970. Ring closure and local conformational deformations of chain molecules. *Macromolecules 3*:178–187.

Greer J. 1985. Molecular structure of the inflammatory protein C5a. *Science 228*:1055–1060.

Greer J. 1990. Comparative modeling methods: Application to the family of mammalian serine proteases. *Proteins Struct Funct Genet 7*:317–334.

Havel TF. 1993. Predicting the structure of the flavodoxin from *Escherichia coli* by homology modeling, distance geometry and molecular dynamics. *Mol Sim 10*:175–210.

Havel TF, Snow ME. 1991. A new method for building protein conformations from sequence alignments with homologues of known structure. *J Mol Biol 217*:1–7.

Hyberts SG, Goldberg MS, Havel TF, Wagner G. 1992. The solution structure of eglin C based on measurements of many NOES and coupling constants and its comparison with X-ray structures. *Protein Sci 1*:736–751.

Isaac VE, Sciavolino P, Abate C. 1995. Multiple amino acids determine DNA binding specificity of the Msx-1 homeodomain. *Biochemistry 34*:7127–7134.

Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein folding recognition. *Nature 358*:86–89.

Kessel M, Gruss P. 1990. Murine developmental control genes. *Science 249*:374–379.

Kissinger CR, Liu B, Martin-Blanco E, Kornberg TB, Pabo CO. 1990. Crystal structure of an engrailed homeodomain–DNA complex at 2.8 Å resolution: A framework for understanding homeodomain–DNA interactions. *Cell 63*:579–590.

Klemm JD, Rould MA, Aurora R, Herr W, Pabo CO. 1994. Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. *Cell 77*:21–32.

Kornberg TB. 1993. Understanding the homeodomain. *J Biol Chem 268*:26813–26816.

Krumlauf R. 1994. *Hox* genes in vertebrate development. *Cell 78*:191–201.

Laughon A. 1991. DNA binding specificity of homeodomains. *Biochemistry 30*:11357–11367.

Leiting B, De Francesco R, Tomei L, Cortese R, Otting G, Wüthrich K. 1993. The three-dimensional NMR-solution structure of the polypeptide fragment 195–286 of the LFB1/HNF1 transcription factor from rat liver comprises a non-classical homeodomain. *EMBO J 12*:1797–1803.

Levitt M. 1992. Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol 226*:507–533.

Lewis PN, Momany FA, Scheraga HA. 1973. Chain reversals in proteins. *Biochim Biophys Acta 303*:211–229.

Li T, Stark MR, Johnson AD, Wolberger C. 1995. Crystal structure of the MATa1/MATα2 homeodomain heterodimer bound to DNA. *Science 270*:262–269.

Odenwald WF, Garbern J, Arnheiter H, Tournier-Las-server E, Lazzarini RA. 1989. The Hox-1.3 homeo box protein is a sequence-specific DNA-binding phosphoprotein. *Genes & Dev 3*:158–172.

Otting G, Qian YQ, Müller M, Affolter M, Gehring W, Wüthrich K. 1988. Secondary structure determination for the Antennapedia homeodomain by nuclear magnetic resonance and evidence for a helix-turn-helix motif. *EMBO J 7*:4305–4309.

Palmer KA, Scheraga HA, Riordan JF, Vallee BL. 1986. A preliminary three-dimensional structure of angiogenin. *Proc Natl Acad Sci USA 83*:1965–1969.

Phillips CL, Vershon AK, Johnson AD, Dahlquist FW. 1991. Secondary structure of the homeo domain of yeast α2 repressor determined by NMR spectroscopy. *Genes & Dev 5*:764–772.

Ponder JW, Richards FM. 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.

Qian YQ, Billeter M, Otting G, Müller M, Gehring WJ, Wüthrich K. 1989. The structure of the *Antennapedia* homeodomain determined by NMR spectroscopy in solution: Comparison with prokaryotic repressors. *Cell 59*:573–580.

Qian YQ, Otting G, Billeter M, Müller M, Gehring WJ, Wüthrich K. 1993. NMR spectroscopy of a DNA complex with the uniformly ^{13}C-labeled *Antennapedia* homeodomain and structure determination of the DNA-bound homeodomain. *J Mol Biol 234*:1070–1083.

Qian YQ, Resendez-Perez D, Gehring W., Wüthrich K. 1994. The des (1–6) Antennapedia homeodomain: Comparison of the NMR solution structure and DNA-binding affinity with the intact Antennapedia homeodomain. *Proc Natl Sci Acad USA 91*:4091–4095.

Rivera-Pomar R, Niessling D, Schmidt-Ott U, Gehring WJ, Jäckle H. 1996. RNA binding and translational suppression by bicoid. *Nature 379*:746–749.

Sali A. 1995. Modelling mutations and homologous proteins. *Curr Opin Biotechnol 6*:437–451.

Sali A, Blundell TL. 1993. Comparative protein modeling by satisfaction of spatial restraints. *J Mol Biol 234*:779–815.

Scheraga HA. 1984. Recent progress in the thermodynamic treatment of protein folding. *Biopolymers* 22:1–14.

Schneuwly S, Kuroiwa A, Baumgartner P, Gehring WJ. 1986. Structural organization and sequence of the homeotic gene *Antennapedia* of *Drosophila melanogaster*. *EMBO J 5*:733–739.

Scott MP, Tamkun JW, Hartzell GW. 1989. The structure and function of the homeodomain. *Biochim Biophys Acta 989*:25–48.

Shang Z, Isaac VE, Li H, Patel L, Catron KM, Curran T, Montelione GT, Abate C. 1994. Design of a "minimAl" homeodomain: The N-terminal arm modulates DNA binding affinity and homeodomain structure. *Proc Natl Acad Sci USA 91*:8373–8377.

Sivaraja M, Botfield MC, Mueller M, Jancso A, Weiss M. 1994. Solution structure of a POU-specific homeodomain: 3D-NMR studies of human B-cell transcription factor Oct-2. *Biochemistry 33*:9845–9855.

Snow ME. 1993. A novel parameterization scheme for energy equations and its use to calculate the structure of protein molecules. *Proteins Struct Funct Genet 15*:183–190.

Song K, Wang Y, Sassoon D. 1992. Expression of Hox-7.1 in myoblasts inhibits terminal differentiation and induces cell transformation. *Nature 360*:477–481.

Srinivasan S, March CJ, Sudarsanam S. 1993. An automated method for modeling proteins on known templates using distance geometry. *Protein Sci 2*:277–289.

Sudarsanam S, March CJ, Srinivasan S. 1994. Homology modeling of divergent proteins. *J Mol Biol 241*:143–149.

Tejero R, Bassolino-Klimas D, Bruccoleri RE, Montelione GT. 1996. Simulated annealing with restrained molecular dynamics using CONGEN: Energy refinement of the NMR solution structures of epidermal and type-α transforming growth factors. *Protein Sci 5*:572–592.

Wang BB, Muller-Immergluck MM, Austin J, Robinson NT, Chisholm A, Kenyon C. 1993. A homeotic gene clutster patterns the anteroposterior body axis of *C. elegans*. *Cell 74*:29–42.

Wolberger C. Homeodomain interactions. 1996. *Curr Opin Struct Biol* 6:62–68.

Wolberger C, Vershon AK, Liu B, Johnson AD, Pabo CO. 1991. Crystal structure of a Matα2 homeodomain–operator complex suggests a general model for homeodomain–DNA interactions. *Cell 67*:517–528.