PROTEINS: Structure, Function, and Genetics 23:301–317 (1995)

A Critical Assessment of Comparative Molecular Modeling of Tertiary Structures of Proteins*

Steven Mosimann, Ron Meleshko, and Michael N.G. James
Medical Research Council of Canada, Group in Protein Structure and Function, Department of Biochemistry, University of Alberta, Edmonton, Alberta T6G 2H7, Canada

ABSTRACT **In spite of the tremendous increase in the rate at which protein structures are being determined, there is still an enormous gap between the numbers of known DNA-derived sequences and the numbers of three-dimensional structures. In order to shed light on the biological functions of the molecules, researchers often resort to comparative molecular modeling. Earlier work has shown that when the sequence alignment is in error, then the comparative model is guaranteed to be wrong. In addition, loops, the sites of insertions and deletions in families of homologous proteins, are exceedingly difficult to model. Thus, many of the current problems in comparative molecular modeling are minor versions of the global protein folding problem. In order to assess objectively the current state of comparative molecular modeling, 13 groups submitted blind predictions of seven different proteins of undisclosed tertiary structure. This assessment shows that where sequence identity between the target and the template structure is high (> 70%), comparative molecular modeling is highly successful. On the other hand, automated modeling techniques and sophisticated energy minimization methods fail to improve upon the starting structures when the sequence identity is low (~30%). Based on these results it appears that insertions and deletions are still major problems. Successfully deducing the correct sequence alignment when the local similarity is low is still difficult. We suggest some minimal testing of submitted coordinates that should be required of authors before papers on comparative molecular modeling are accepted for publication in journals.** © 1995 Wiley-Liss, Inc.

Key words: molecular model, comparative model, homology model, structure prediction, calculated structure

INTRODUCTION

Once a protein's sequence has been determined and it has been found to be a new member of a structurally characterized protein family, it is relatively straightforward to build a molecular model of the protein using a set of simple guidelines.[1,2] Presently, there are several commercial and public domain computer programs that have been developed for modeling; these programs remove much of the tedium from the process. There are numerous reasons for constructing comparative molecular models of proteins. The molecular model may explain the structural basis of existing experimental results and can provide one with structural information on which further experiments can be planned, executed, and evaluated. Site-specific mutations of the gene coding for the specific protein can provide important data regarding the protein's function. Perhaps, some of the most revealing experiments are those designed to predict and to probe the molecular reasons for an enzyme's specificity.[3] On a more practical note, the molecular model can sometimes be used successfully to determine phases for a crystal structure determination using the method of molecular replacement.[4] The more spectacular uses, however, are typified by the recent successful application of comparative molecular modeling for identifying new classes of lead compounds in antimalarial drug development.[5]

An example of the successful prediction of an enzyme's specificity from comparative molecular modeling is that for granzyme B (CCP1), a serine proteinase from cytotoxic T lymphocytes.[3] A molecular model of CCP1 (48% identical to rat mast cell proteinase II) showed that an arginine at position 226 would occupy the S_1 specificity pocket, thereby suggesting a P_1 specificity for an aspartate or glutamate residue. Subsequent synthesis and testing of a series of substrates differing in the nature of the P1 residue confirmed the aspartate specificity of CCP1.[6] The P_1 specificity of CCP1 has recently been altered by site-specific mutagenesis of the residue at

*This assessment does not indicate that any one particular modeling group or modeling technique is superior to any other. We do not believe that comparative molecular models can be ranked using a single or even several numeric indicators. As such, claims that particular modeling techniques are superior based upon the results herein are not justifiable, in our opinion.

Received March 30, 1995; revision accepted June 20, 1995.
Address reprint requests to Michael N.G. James, Medical Research Council of Canada, Group in Protein Structure and Function, Department of Biochemistry, University of Alberta, Edmonton, Alberta T6G 2H7, Canada.

position 226 to a chymotryptic-like specificity.[7] The successes of these experiments indicate that the active site and the S_1 specificity pocket of the comparative molecular model of CCP1 must be close to the true structure.

In spite of an ever-increasing number of such studies that have demonstrated the broad utility of comparative modeling, there have only been a few for which the modeling process has been assessed objectively. It was not until quite recently[8] that the first-ever constructed comparative model, that of bovine α-lactalbumin[9,10] based on the structure of hen egg white lysozyme[11] was evaluated. The α-lactalbumin model was constructed from brass Kendrew-model components (scale, 2 cm = 1 Å) and measured manually. Remarkably, the agreement between the model structures, assessed by comparing the C^α coordinates with the experimentally determined α-lactalbumin C^α coordinates, is extraordinarily good (rms deviation based on C^αs of residues 1–95 is 1.1 Å). The C-terminal ~30 residues of α-lactalbumin are less well ordered and agree less well with those homologous residues of lysozyme.

Serine proteinases have been a popular target for comparative molecular modeling. The three-dimensional structures of the two serine proteinases, porcine elastase[12] and bovine α-chymotrypsin,[13] were used to construct a comparative molecular model of the distantly related bacterial serine proteinase, α-lytic protease.[14] Evaluation of this model[15] was based on the pulished description of the model and the topographical superposition of α-lytic protease[16] with porcine elastase.[12] It was clear from that analysis that when two proteins are so distantly related (~18% identity in the structurally aligned sequences[17]), it is extremely difficult, if not impossible, to determine a correct overall alignment of the sequences. It was also apparent that major segments of the polypeptide chain were folded completely differently in α-lytic protease and elastase in spite of correct local alignment. The cores of the two enzymes are similar; 108 C^α atoms of both enzymes have an rms deviation of 2.08 Å. This structural similarity, however, is limited mainly to the strands that support the catalytic residues or form the substrate binding sites. The limited sequence identity between the two enzymes precluded the correct alignment even for several of the core β strands. "Where the alignment is in error the model is guaranteed to be wrong" (comparative modeling maxim attributed to R. Read[18]).

Shorty after the complete amino-acid sequence of thrombin became available, a molecular model of the catalytic B chain of bovine thrombin[19] was constructed, based on the known structure of α-chymotrypsin. This model has not been evaluated in light of the known thrombin structure.[20–22] A critical evaluation of molecular models of the bacterial trypsin from *S. griseus*, SGT, showed that local structure is conserved among members of protein families more strongly than global structure. As a result, one often observes relative shifts of secondary structural elements.[18,23,24] The analysis of SGT also showed that the predictions of nonhomologous regions in proteins are much less reliable than the predictions of highly conserved regions. The reliability of a model is especially problematic if one is interested in the unique substrate specificity of an enzyme (e.g., renin[25]), because these are the regions for which the model is likely to be most inaccurate. The several models of human renin were evaluated only qualitatively and so far not systematically in light of the crystal structure of renin.[26]

There are now many three-dimensional molecular models that have been constructed for a variety of proteins. This "competition" was designed to assess objectively just where comparative molecular modeling stands in terms of being an art or a science. In this paper we present an analysis of the results of predictions by 13 different groups for 7 different proteins (Table I) whose tertiary structures were in the process of being determined by X-ray crystallographic techniques. The 13 groups (Table II) submitted a total of 43 separate coordinate sets for the 7 protein structures. Some proteins received a lot of attention [eosinophil-derived neurotoxin (EDN) had 11 predictions; cellular retinoic acid binding protein I (CRABPI) had 10 predictions] whereas others attracted only a few submitted predictions P450 (2) and HFD (1). One additional prediction was made for an eighth protein, a dihydrofolate reductase (Table I). As the three-dimensional structure was not determined in time, this protein could not be included in our evaluation.

Table I lists the 7 protein structures that comprise the unknowns. The crystal structures available for these 7 proteins are of medium (2.8 to 2.9 Å) to high (1.8 Å) resolution and all had been at least partially refined at the close of the competition, October 31, 1994. The quality of each of the experimental structures was evaluated by the program PROCHECK[45] to ensure that all were of acceptable accuracy to evaluate the predictions.

METHODS

Protein structures, suitable for comparative molecular modeling, not yet published and therefore unavailable to the public, were solicited by the conference organizers. These proteins were then made available to contestants via anonymous ftp. Participants were provided with sequence information, a list of relevant references, if available, and an indication of the state of the experimental work. Predictions were submitted for evaluation in Protein Data Bank (PDB[29]) format via e-mail to an account (homology@biochem.ualberta.ca) created expressly for this purpose. The deadline for all submissions was no later than October 31, 1994; earlier deadlines ap-

TABLE I. The Target X-Ray Crystallographic Structures for Comparative Molecular Modeling

Name	Protein type	Investigators	Template*; identity†	Resolution (Å)	R-Factor‡ (Oct. 94)	Number of predictions submitted	Number of labs
NM23§	NDP kinase	R. Williams	1NDL 77%	2.8	0.24	7	6
E5.2	Immunoglobulin domain	B. Fields R. Poljak	1FAI 76%	1.9	0.19	3	3
HPR	Phosphocarrier protein	U. Pieper O. Herzberg	2HPR 42%	1.8	0.17	9	8
CRABPI**[27]	Lipocalin	G. Kleywegt T.A. Jones	2HMB 41%	2.9	0.25	10	6
HFD	Ferredoxin	M. Shoham et al.	1FXA 40%	1.9	0.17	1	1
EDN	RNase	S. Mosimann M. James	6RSA 35%	1.8	0.17	11	7
P450[28]	Heme protein	J. Cupp-Vickery T. Poulos	1CPT 22%	2.1	0.19	2	2
HVDHFR	Dihydrofolate reductase		Not completed in time			1	1

*Template structure refers to the structure that provided the initial coordinates for the comparative molecular models. The templates are referenced by their Protein Data Bank[29] accession codes.

†The reported percentage identity is based upon the number of identical residues that can be superimposed within 3.8 Å in the C^{α} atom least-squares superposition of the target and template structures.

‡The conventional R-factor is

$$R = \frac{\sum||F_o| - |F_c||}{\sum|F_o|},$$

where $|F_o|$ and $|F_c|$ are the observed and calculated structure factor amplitudes, respectively.

§NM23 is a hexamer of identical subunits. The rmsds for the pairwise comparison of the NM23 subunits range from 0.30 to 0.57 Å for the main chain atoms (R. Williams, personal communication). All submitted comparative molecular models consist of a single NM23 subunit. The N subunit of NM23 was chosen as the representative NM23 subunit for all comparisons as it is the only subunit having complete coordinates for the C-terminal residue, Glu-152.

**The CRABPI crystal structure contained two copies of the protein. The coordinates were refined using noncrystallographic symmetry constraints[27] and the two copies are identical.

plied to those structures that were likely to become public prior to this date. In practice, a few submissions were received after the final deadline, but in all cases the blind nature of the process was maintained. After the deadlines had passed, the coordinates of the crystallographically determined structures were provided by the groups that had solved the structures.

When referring to a structure, we will use "prediction" to indicate a structure obtained via comparative modeling techniques, "experimental" or "target" to refer to a structure obtained via X-ray crystallographic techniques, and "parent" or "template" to refer to a three-dimensional coordinate set that was used as the starting point for a comparative modeling solution.

The submitted PDB files for the predictions, along with those for the experimental structures, were first evaluated by PROCHECK (v 3.0.1 or v 3.2)[45] in order to generate detailed stereochemical information about the structures† and to ensure consistent naming of side-chain atoms. In addition to PROCHECK, RMS6 (provided to us by the conference organizers and enhanced in-house), O[46], BBDEP94[47], and an assortment of AWK scripts[48] were used to generate analyses of the predictions. PROCHECK produces a number of text files containing extensive structural information. In particular, the .out file contains information about the secondary structure and the ϕ,ψ distributions; the .rin (residue information) file contains information about side-chain conformations (χ_1, χ_2), peptide bond planarity (ω), and solvent accessibility, and the .nb file has information about nonbonded contacts within a structure. In some instances, the desired information was obtained directly from these files, but in the majority of the cases, AWK scripts were written to extract and transform the raw data generated by PROCHECK.

RMS6 was used to superimpose molecular struc-

†Some preprocessing of the submitted coordinates had to be performed. This included checking for inadvertent insertions and deletions, for erroneous sequence numbering, for any required formatting changes, for the renaming of some atoms, and for ensuring that all occupancies were nonzero.

TABLE II. The Participating Comparative Molecular Model Builders

Laboratory	Institution	Code	Number of structures predicted	Number of predictions submitted	References
T. Cardozo M. Totrov R. Abagyan	New York University, USA	aba	5	5	30
M. Bolger S. Basu	USC Los Angeles, USA	bol	1	1	31,32
B. Church (EDN) D. Kitson (HPR)	Biosym, USA/UK	bio	2	2	33
M. Delarue P. Koehl	Institut Pasteur, France U. Strasbourg, France	koe	3	4	34
J. Pedersen R. Samudrala H-B. Zhou R. Luo K. Fidelis J. Moult	CARB and Lawrence Livermore National Lab. (KF), USA	mou	3	3	35
Oxford Molecular	Oxford Molecular, USA/UK	oxm	1	1	36
A. Sali L. Potterton F. Yuan H. van Vlijmen M. Karplus	Harvard University, USA	sal	3	4	37
M. Saqi	Glaxo Group Research, UK	saq	1	2	38,39
D. Mosenkis	Tripos, USA	tri	1	1	40
M. Vihinen	Karolinska Institute, Sweden	vih	2	2	33,41
C. Vinals P. Briffeuil E. Feytmans	Facultés Universitaires ND de la Paix, Belgium	vin	2	6	42
G. Vriend	EMBL Heidelberg, Germany	vri	3	3	43
R. Harrison D. Chatterji I. Weber	Thomas Jefferson U., USA	web	7	9	44

tures and to calculate the rmsds[‡] using various atomic pairings of the experimental, template,[§] and predicted structures. The superpositions used atoms (C^{α} or all atoms) of specified pairs of residues and the techniques of Kabsch[49] and McLachlan.[50] In addition to using all residues, residues were selected on the basis of the magnitude of the *B*-factor (Å^2) and on secondary structural characteristics. For the selection of atoms on the basis of *B*-factor, those atoms with *B* less than ~1.2 times the mean *B* of the experimental structure were selected. For the secondary structure comparisons, those residues in the experimental structures adopting α-helical or β-strand conformations (Kabsch and Sander classification[51] produced by PROCHECK) were chosen. Other comparison criteria were obtained by preprocessing the input PDB files using AWK scripts.

Superpositions based on structural similarity were obtained using O (v 5.9 and v 5.10). This was done using the lsq_implict command with the default cutoff of 3.8 Å.** The structural superpositions of the experimental and template structures were used to obtain a measure of the identity of the two proteins. The percent identity reported in Table I was obtained from the number of identical residues having C^{α} atoms within 3.8 Å in the superposition of

‡ $\text{Rmsd} = \left(\sum_{N} \frac{d(a_i,b_i)^2}{N}\right)^{1/2}$

where *N* is the number of pairs of atoms, a_i and b_i are the two in pair *i*, and $d(\ )$ is the Euclidean distance operator.

§In the case of more than one potential template structure, the most commonly used template was selected for inclusion in Table IV.

**For P450 and E5.2 initial structural superpositions had to be performed manually.

††The only exception to this was P450 where the percentage identity was provided by T. Poulos (personal communication).

the parent and target structures.[††] O was also used to examine the predictions graphically and to compare the superimposed structures from the predictions with the experimental structures.

The program BBDEP94, coupled with AWK scripts, was used to evaluate the χ_1 conformations of both the target and predicted structures. The value reported is the percentage of χ_1 angles in the structure that adopt the preferred ϕ,ψ dependent conformation of χ_1 as determined by the September 1994 BBDEP94 database (268 chains from 253 proteins solved at 2.0 Å resolution or better).

RESULTS

The experimental structures are from diverse functional and structural protein families. NM23 and HPR participate in transphosphorylation reactions; P450[28] and HFD undergo electron transfer reactions and bind prosthetic groups involving Fe^{2+}; E5.2 and CRABPI[27] bind ligands and EDN is an endoribonuclease. The two immunoglobulin domains of E5.2 are all β-structures; the other experimental structures are a mixture of α-helix and β-sheet. Each of the submitted coordinate sets is of high quality and their associated stereochemical parameters, as evaluated by PROCHECK,[45] are consistent with well-refined crystal structures.

Table III presents a brief summary of the stereochemical quality of all the experimental and predicted protein structures. The ϕ,ψ distributions of the Ramachandran plots are given as percentages of the total number of nonglycine, nonproline residues found in the four different categories (PROCHECK definition). The peptide bond planarity is given as an rmsd from 180° (or 0° for *cis*-peptide bonds) in the ω angle for the four atoms (C^{α}_i, C_i, N_{i+1}, C^{α}_{i+1}) of all the peptide bonds in the structures. The experimental structures all have profiles in the Ramachandran plot expected for well-determined crystal structures. There are only four residues in the so-called "disallowed regions" and these residues are parts of loop structures. Two of the residues (P450: Glu-161 and P450:Ala-163) in the "disallowed regions" have average *B*-factors significantly larger than the average *B*-factor of the structure and are in relatively poorly determined regions of the experimental structure. The ϕ,ψ angles for NM23:Ile-116 (42.34°, −45.78°) and E5.2:Thr-51 (77.34°, −61.63°) fall outside of the left-handed α-helix region. Peptide bond planarity rmsds for the experimental structures are all well within the expected range as defined by PROCHECK.[45]

The comparative molecular models have a wide range of Ramachandran plot profiles. Some of the predicted structures have a significantly higher percentage of residues in the "most favored region" than the experimental structures (i.e., NM23 88.1% vs. NM23sal with 96.8%) whereas others have significantly lower percentages (i.e., CRABPI 82.6% vs. the fully automated CRABPIvin2 with 51.6%, or EDN 88.1% vs. EDNweb with 59.5%). The predicted and experimental protein structures have comparable numbers of residues in the "generously allowed" and "disallowed" regions of the Ramachandran plot except in the case of EDN. The comparative molecular models sharing the highest-degree of amino acid sequence identity with their template structure(s) have Ramachandran plot profiles that are most similar to those of the experimental structure.

The submitted models also have a wide range of ω angle rmsds from planarity. For 24 of the 43 predicted structures the ω angle rmsd from planarity is *greater than* 10.0°. Some of the submitted predictions (NM23web1; HPRaba; CRABPIweb1, web2; HFDweb; EDNmou, web; P450web) have rmsds in ω greater than 20° from *trans* or *cis* planar peptide bonds. In comparison, the rmsd of peptide bond planarity for the seven experimental structures ranges between 1.2° (NM23 and P450) and 4.3° (HPR). The typical rmsd from planarity of peptide bonds used by the structure verification program PROCHECK is 6.0 ± 3.0°. Some of the laboratories consistently have large departures from ω bond planarity (aba, web). This suggests that the corresponding force constants in energy minimization routines, or their equivalent, are inappropriate thereby allowing such large deviations.

The χ_1 rotamers of the nonglycine, nonalanine residues in each structure have been compared to the preferred χ_1 rotamer for each amino-acid type and the corresponding set of ϕ,ψ torsion angles[47] (see Table III). The experimental structures have from 65% (NM23 and HFD) to 74% (HPR and EDN) of their side chain χ_1 angles in the preferred conformation in the rotamer library. The seven comparative molecular models of NM23 (77% sequence identity to 1NDL, accession code for PDB[29]) have a median value of 66%, comparable to that of the experimental structure. E5.2 has 76% sequence identity to its template structure. The predicted models of E5.2 all have χ_1 distributions having considerably fewer (from 11 to 16% less) residues in the most favorable χ_1 range than has the experimental structure. On the other hand, the 11 predicted structures of EDN (35% identical to 6RSA) have an average of 54% of the residues with χ_1 in the preferred conformation. This is considerably less than the value of 74% for the experimental structure of EDN (Table III). As the sequence identity of the experimental and template structures decreases to less than 40%, some of the predictions have χ_1 distributions that have from 24% (EDNweb) to 32% (EDNvin3) fewer residues having the preferred conformational angles relative to the experimental structures.

Table III also includes a summary of the distribution of intramolecular nonbonded contacts for all structures. P450 is the only experimental structure having a nonbonded contact less than 2.5 Å in

TABLE III. Selected Stereochemical Properties for the Experimental X-Ray Crystallographic Structures and the Comparative Molecular Models

	Predicted structures							
NM23 (residues: 148 total; 126 non-Gly, non-Pro; 126 with χ_1)								
	targ	aba	koe	sal	vih	vri	web1	web2
Most favored regions*	88.1	92.8	83.3	96.8	95.0	91.2	86.5	84.1
Additional allowed regions*	9.5	6.4	15.9	2.4	4.2	7.2	11.1	14.3
Generously allowed regions*	1.6	0.8	0.8	0.0	0.0	0.8	0.8	0.8
Disallowed regions*	0.8	0.0	0.0	0.8	0.8	0.8	1.6	0.8
Peptide ω angle, rmsd (°)†	1.4	8.8	5.8	2.9	6.8	5.9	20.3	15.7
Preferred χ_1 (%)‡	58	72	71	75	64	68	56	59
No. D-amino acids	0	0	0	0	0	0	0	0
No. *cis*-peptides	0	0	0	0	0	1	0	0
No. nonbonded contacts§								
< 2.5 Å	0	0	0	0	15	0	0	0
2.5 to 3.0 Å	77	68	17	81	68	47	76	79
3.0 to 3.5 Å	96	210	273	210	195	207	190	174
E5.2 (residues: 228 total; 196 non-Gly, non-Pro; 193 with χ_1)								
	targ	aba	bol	oxm				
Most favored regions*	86.7	80.8	72.9	83.8				
Additional allowed regions*	12.8	14.6	23.4	13.1				
Generously allowed regions*	0.0	2.0	1.0	0.5				
Disallowed regions*	0.5	2.5	2.6	2.5				
Peptide ω angle, rmsd (°)†	2.1	16.4	7.3	7.3				
Preferred χ_1 (%)‡	69	53	56	58				
No. D-amino acids	0	0	0	3				
No. *cis*-peptides	2	2	2	1				
No. nonbonded contacts§								
< 2.5 Å	0	4	0	0				
2.5 to 3.0 Å	67	60	45	35				
3.0 to 3.5 Å	280	277	322	299				
HPR (residues: 88 total; 81 non-Gly, non-Pro; 69 with χ_1)										
	targ	aba	bio	koe1	koe2	mou	tri	vih	vri	web
Most favored regions*	90.1	82.7	85.2	91.1	92.2	93.7	85.2	91.1	88.8	88.9
Additional allowed regions*	8.6	16.0	12.3	8.9	6.5	5.1	13.6	8.9	11.3	11.1
Generously allowed regions*	1.2	1.2	1.2	0.0	1.3	1.3	1.2	0.0	0.0	0.0
Disallowed regions*	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Peptide ω angle, rmsd (°)†	3.1	19.4	6.9	2.6	3.0	9.1	2.7	2.6	6.1	12.5
Preferred χ_1 (%)‡	74	64	55	73	73	58	58	52	68	55
No. D-amino acids	0	0	0	0	0	0	0	0	0	0
No. *cis*-peptides	0	0	0	0	0	0	0	0	0	0
No. nonbonded contacts§										
< 2.5 Å	0	0	0	0	0	0	0	8	10	0
2.5 to 3.0 Å	41	38	6	34	35	19	69	46	49	47
3.0 to 3.5 Å	118	114	124	117	109	132	107	110	104	104
CRABPI (residues: 136 total; 122 non-Gly, non-Pro; 116 with χ_1)											
	targ	aba	mou1	mou2	sal	vin1	vin2**	vin3**	vri	web1	web2
Most favored regions*	82.6	80.3	86.9	82.8	87.7	73.8	63.9	51.6	91.5	72.1	73.8
Additional allowed regions*	15.7	16.4	9.0	13.9	11.5	23.8	27.9	35.2	6.0	24.6	23.8
Generously allowed regions*	1.7	1.6	0.8	0.8	0.8	0.8	1.6	7.4	0.9	1.6	1.6
Disallowed regions*	0	1.6	3.3	2.5	0.0	1.6	6.6	5.7	1.7	1.6	0.8
Peptide ω angle, rmsd (°)†	1.5	11.4	12.8	11.8	2.1	11.6	10.7	11.9	5.2	26.4	27.9
Preferred χ_1 (%)‡	66	63	55	59	66	53	42	38	63	55	59
No. D-amino acids	0	0	0	0	0	4	4	0	1	1	0
No. *cis*-peptides	0	0	0	0	0	0	0	0	0	0	0
No. nonbonded contacts§											
< 2.5 Å	0	2	0	0	0	1	1	1	0	0	0
2.5 to 3.0 Å	39	34	26	23	80	12	16	18	24	43	37
3.0 to 3.5 Å	174	160	186	175	146	172	174	184	143	174	172

(Continued)

TABLE III. Selected Stereochemical Properties for the Experimental X-Ray Crystallographic Structures and the Comparative Molecular Models *(continued)*

HFD (residues: 128 total; 116 Non-Gly, non-Pro; 107 with χ_1)

	targ	web
Most favored regions*	87.9	80.8
Additional allowed regions*	12.1	14.6
Generously allowed regions*	0.0	2.0
Disallowed regions*	0.0	2.5
Peptide ω angle, rmsd (°)†	4.3	35.6
Preferred χ_1 (%)‡	65	49
No. D-amino acids	0	8
No. *cis*-peptides	0	0
No. nonbonded contacts§		
< 2.5 Å	0	0
2.5 to 3.0 Å	47	37
3.0 to 3.5 Å	168	118

EDN (residues: 134 total; 118 non-Gly, non-Pro; 124 with χ_1)

	targ	bio	koe	mou	sal1	sal2	saq1	saq2	vin1	vin2**	vin3**	web
Most favored regions*	88.1	59.7	83.8	74.6	87.3	85.6	82.8	82.9	72.0	55.1	69.5	59.5
Additional allowed regions*	11.9	26.9	13.7	19.5	11.0	12.7	12.9	13.7	22.0	32.2	23.7	28.4
Generously allowed regions*	0.0	6.7	0.0	4.2	1.7	1.7	2.6	2.6	3.4	7.6	3.4	7.8
Disallowed regions*	0.0	6.7	2.6	1.7	0.0	0.0	1.7	0.9	2.5	5.1	3.4	4.3
Peptide ω angle, rmsd (°)†	3.0	14.7	6.8	21.0	3.3	4.1	19.4	11.9	11.2	12.0	11.0	38.1
Preferred χ_1 (%)‡	74	52	59	54	66	65	47	67	52	44	42	50
No. D-amino acids	0	0	0	2	0	0	1	0	2	12	12	9
No. *cis*-peptides	0	0	0	2	0	0	2	0	4	0	0	2
No. nonbonded contacts§												
< 2.5 Å	0	0	0	0	2	2	21	3	0	0	0	0
2.5 to 3.0 Å	53	30	16	63	115	107	61	72	20	21	12	65
3.0 to 3.5 Å	158	170	212	179	144	155	168	163	158	186	185	125

P450 (residues: 403 total; 348 non-Gly, non-Pro; 336 with χ_1)

	targ	aba	web
Most favored regions*	89.7	84.2	76.1
Additional allowed regions*	9.8	14.7	19.8
Generously allowed regions*	0.0	0.6	2.0
Disallowed regions*	0.6	0.6	2.0
Peptide ω angle, rmsd (°)†	1.2	11.2	29.0
Preferred χ_1 (%)‡	69	55	53
No. D-amino acids	0	0	5
No. *cis*-peptides	1	1	1
No. nonbonded contacts§			
< 2.5 Å	1	0	0
2.5 to 3.0 Å	194	236	209
3.0 to 3.5 Å	543	537	481

*The percent of residues falling into the four regions of the Ramachandran plot (defined by PROCHECK[45]) for each experimental structure and the comparative molecular models.

†The peptide bond planarity is reported in terms of the ω torsion angle (C^{α}_i, C_i, N_{i+1}, C^{α}_{i+1}) rmsd from 180° (*trans*-peptide) or 0° (*cis*-peptide).

‡The χ_1 torsion angles of each structure were compared to the backbone dependent rotamer library of Dunbrack and Karplus.[47] The value reported is the percentage of residues of a given type that adopt a χ_1 torsion angle corresponding to the preferred χ_1 rotamer, given its ϕ,ψ angles.

§Nonbonded contacts do not include potential hydrogen bonds between 2.5 and 3.5 Å.

**These are fully automated predictions using the commercially available Biosym software.

length. This short contact (Asp-160 O–Glu-161 O) is part of the poorly determined, high *B*-factor loop mentioned previously. In contrast, 12 of the predicted structures have nonbonded contacts less than 2.5 Å. Four of these (NM23vih; HPRvih, vri; EDN-saq1) have more than 9 such contacts per 100 residues. Apart from these four structures for which this distribution should cause concern, the numbers of nonbonded contacts in the three ranges chosen are of similar magnitude in the experimental and predicted structures.

The E5.2 and P450 experimental structures and the parent structures for the E5.2, P450, and EDN models contain *cis*-peptide bonds (Table III). The predicted structures accurately predict 5 of the 8 conserved *cis*-peptides and 7 of the 11 EDN models correctly predict its lack of *cis*-peptides. Table III includes the number of *cis*-peptides and incorrect chiral centers for each structure.

A very disturbing feature of the submitted comparative model coordinates is the presence of D-chirality for some of the C^{α} atoms and the wrong chi-

rality for some of the β-branched threonine and isoleucine residues. Of the 43 predicted structures, 14 contained one or more atoms with the alternate enantiomorph. Some of these structures also had major departures from peptide bond planarity (e.g., EDNweb, Table III) suggesting that the energy minimization routines were insufficiently or inappropriately constrained. It is unfortunate that some coordinate sets were submitted containing D-amino acid residues as such glaring errors can easily be detected and corrected.

Table IV and Figure 1a–g summarize the results of the various least-squares superpositions between the comparative molecular models and their respective X-ray crystallographic structures. The rmsds for the all atom superpositions range from 1.39 Å (NM23sal) to 10.53 Å (HFDweb). The equivalent C^{α} rmsds vary between 0.53 Å (NM23sal) and 9.94 Å (HFDweb). The magnitudes of the all atom and C^{α} rmsds between the target and predicted protein structures vary according to (1) the percent amino acid sequence identity in the alignment of the target and the template sequences, (2) the size and number of insertions or deletions in the structures, and (3) the modeling groups and their software. Undoubtedly there are other factors that contribute to the magnitude of these rmsds, including the differences between the restrained crystallographic refinement and energy minimization target functions. Also listed in Table IV is the percentage amino acid sequence identity for the structure-based alignment of the target structure and the template structure used to construct the model. Where multiple template structures are available (i.e., NM23: 1NDL, 1NDC, 1NDK, and 1NDP, accession codes for PDB[29]), Table IV lists the one with the highest percent sequence identity.

The all atom and C^{α} rmsds are lowest for the structures with the greatest sequence identity and the rmsds are highest for those structures with the least sequence identity. In Table IV, however, the comparisons of E5.2 and HFD with their respective models do not appear to follow this trend. In the case of E5.2, the quoted rmsds are the result of treating the light and heavy chain fragments as an intact unit. In fact, there are differences in the subunit packing between the light and heavy chains in the experimental and template structure that give rise to inflated rmsds. These packing differences are detected when the light and heavy chains of E5.2 are least-squares superimposed on their targets separately. The rmsds for C^{α} superpositions are 0.58 and 0.80 Å for the light and heavy chains, respectively. HFD has a 29 amino acid insertion when compared to the template structure (M. Shoham, personal communication). The single prediction of HFD has an N-terminal sequence misalignment of the first 36 residues leading to the excessively high rmsds. When these two exceptions are taken into account, the magnitude of the rmsds between predicted and experimental structures does vary inversely with the percent sequence identity between the target and template structures.

The size and number of insertions and/or deletions in the target structure also affect the magnitudes of the rmsds. NM23, HPR, and E5.2 (when treating the subunits independently) have no large insertions or deletions and therefore exhibit the lowest rmsds (Table IV). CRABPI has three small insertions totaling six residues and a resulting intermediate rmsd, whereas HFD, EDN, and P450 have large insertions relative to their parent structures and have the largest rmsds. In the superpositions shown in Figure 1e–g, several regions within each prediction show considerable structural variation. The segments showing poorest agreement are the loop regions that are the sites of insertions and deletions between the target and template structures.

Table IV includes the rmsds for the least-squares superpositions of subsets of atoms in the target structures and the comparative molecular models. There are two questions being addressed in these comparisons of coordinate subsets. Are the well-defined regions of the crystal structures (i.e., low *B*-factors) more accurately predicted than ill-defined regions (high *B*-factor)? Are residues in secondary structural units (α-helices and β-sheets) more accurately predicted than those in the loop regions? In general, the results in Table IV show that predictions seem to be more reliable for regions of the protein that are well defined (low *B*-factor) or that involve residues in regular secondary structural units. However, these apparent improvements in predictability may simply reflect the reduced subset of atoms used in the least-squares minimizations. Examination of Figure 1d and f shows that there are segments of secondary structure having small but concerted differences in position between the experimental and predicted structures. An example of these concerted differences involves the last 3 strands of β-sheet in CRABPI (Fig. 1d). It is not yet possible to move the predicted structure sufficiently from the initial template position to account for such differences in relative positions of secondary structural features.

The superpositions of the C^{α}-traces for the experimental structures and the predicted structures (Figs. 1a–g) show clearly that the positions of the secondary structural regions of the comparative models agree better with the equivalent regions in the experimental structures than do the loop regions (Fig. 1d and f). It is not possible to make such inferences simply by referring to the table of rmsds since this single number only gives an overall estimate. In practice, the *B*-factor threshold and the regular secondary structure requirements tend to exclude the same residues from the superpositions, as the loop structures connecting secondary structural units are

TABLE IV. Least-Squares Superpositions of the Experimental Structures, Template Structures, and the Comparative Molecular Models*

	No restrictions		3.8 Å cutoff	Helix and sheet†		*B*-factor < cutoff‡	
	All Atoms	C^{α}	C^{α}	All atoms	C^{α}	All atoms	C^{α}
NM23						$B < 22.0$ Å^2	
1NDL (77%)§			0.54 (148)				
aba	1.45 (1180)	0.58 (148)	0.58 (148)	1.36 (773)	0.46 (96)	1.11 (743)	0.50 (106)
web1	2.36 (1186)	1.14 (148)	0.99 (146)	1.59 (779)	0.59 (96)	1.78 (743)	0.59 (106)
web2	2.36 (1186)	1.16 (148)	0.99 (146)	1.57 (779)	0.59 (96)	1.75 (743)	0.56 (106)
koe	2.33 (1176)	1.44 (148)	1.20 (142)	1.99 (771)	1.07 (96)	1.67 (734)	1.01 (106)
sal	1.39 (1187)	0.53 (148)	0.53 (148)	1.23 (779)	0.42 (96)	1.08 (741)	0.43 (105)
vih	1.93 (1122)	0.97 (142)	0.83 (140)	1.44 (774)	0.47 (96)	1.64 (721)	0.48 (105)
vri	2.40 (1178)	1.18 (147)	0.88 (146)	1.38 (779)	0.47 (96)	1.62 (743)	0.49 (106)
E5.2						$B < 35.0$ Å^2	
1FAI (76%)§			0.82 (220)				
aba	2.44 (1759)	1.47 (228)	1.18 (223)	1.83 (1121)	1.09 (143)	2.35 (1081)	1.45 (154)
bol	2.41 (1761)	1.70 (228)	1.01 (219)	1.52 (1123)	1.07 (143)	2.47 (1081)	1.82 (154)
oxm	2.15 (1761)	1.35 (228)	1.01 (223)	1.48 (1123)	0.87 (143)	2.07 (1081)	1.41 (154)
HPR						$B < 21.0$ Å[d]	
2HPR (42%)§			0.80 (87)				
aba	1.58 (642)	1.05 (88)	0.76 (87)	1.36 (476)	0.68 (65)	0.77 (319)	0.69 (66)
bio	1.74 (644)	1.21 (88)	0.92 (87)	1.46 (476)	0.77 (65)	0.80 (319)	0.73 (66)
koe1	1.71 (616)	0.79 (86)	0.79 (86)	1.68 (465)	0.72 (65)	0.87 (318)	0.73 (66)
koe2	1.89 (601)	1.27 (84)	1.28 (84)	1.81 (465)	1.18 (65)	1.15 (313)	1.09 (64)
mou	1.76 (644)	1.16 (88)	0.93 (87)	1.51 (476)	0.80 (65)	0.88 (319)	0.79 (66)
tri	1.61 (644)	1.06 (88)	0.81 (87)	1.30 (476)	0.64 (65)	0.79 (319)	0.64 (66)
vih	1.69 (624)	0.79 (86)	0.79 (86)	1.69 (471)	0.72 (65)	0.86 (317)	0.73 (66)
vri	1.48 (636)	0.80 (87)	0.80 (87)	1.26 (473)	0.66 (65)	0.69 (316)	0.62 (66)
web	1.65 (643)	1.06 (88)	0.79 (87)	1.41 (476)	0.66 (65)	0.74 (319)	0.64 (66)
CRABPI						$B < 45.0$ Å^2	
2HMB (41%)§			1.41 (122)				
abe	3.52 (1085)	2.87 (136)	1.42 (122)	3.08 (890)	2.14 (108)	3.06 (639)	2.18 (88)
mou1	2.62 (1087)	2.01 (136)	1.34 (127)	2.33 (892)	1.63 (108)	2.18 (639)	1.89 (88)
mou2	2.64 (1087)	2.07 (136)	1.37 (127)	2.35 (892)	1.69 (108)	2.20 (639)	1.93 (88)
sal	2.84 (1084)	2.16 (136)	1.38 (124)	2.45 (889)	1.60 (108)	2.33 (636)	1.90 (88)
vin1	4.24 (1087)	3.32 (136)	1.68 (121)	4.13 (892)	3.19 (108)	3.97 (639)	3.20 (88)
vin2**	4.42 (1087)	3.72 (136)	1.98 (113)	4.18 (892)	3.57 (108)	4.17 (639)	3.44 (88)
vin3**	4.37 (1087)	3.48 (136)	1.86 (113)	4.36 (892)	3.37 (108)	4.22 (639)	3.35 (88)
vri	2.96 (1045)	2.15 (130)	1.40 (123)	2.87 (871)	1.86 (105)	2.84 (623)	2.09 (86)
web1	2.97 (1087)	2.12 (136)	1.64 (132)	2.95 (892)	2.01 (108)	2.80 (639)	2.11 (88)
web2	3.01 (1087)	2.16 (136)	1.70 (131)	2.99 (892)	2.05 (108)	2.81 (639)	2.09 (88)
HFD						$B < 8.0$ Å^2	
1FXA (40%)§			1.27 (90)				
web	10.53 (921)	9.94 (118)	1.68 (93)	8.94 (564)	8.74 (73)	5.85 (533)	5.33 (77)
EDN						$B < 20.0$ Å^2	
6RSA (35%)§			1.29 (106)				
bio	6.18 (1078)	5.15 (134)	1.55 (97)	3.36 (695)	2.03 (85)	4.87 (792)	3.84 (109)
koe	6.13 (1047)	4.96 (131)	1.35 (94)	3.70 (688)	2.35 (85)	4.87 (771)	4.00 (107)
mou	5.45 (1079)	4.53 (134)	1.38 (96)	3.46 (696)	2.35 (85)	4.53 (792)	3.66 (109)
sal1	6.02 (1079)	4.76 (134)	1.38 (108)	5.27 (696)	3.98 (85)	5.80 (714)	4.68 (100)
sal2	5.73 (1074)	4.61 (134)	1.44 (111)	5.20 (696)	3.99 (85)	5.56 (727)	4.59 (102)
saq1	4.47 (1047)	3.03 (130)	1.31 (105)	3.55 (696)	2.17 (85)	3.67 (767)	2.59 (106)
saq2	5.03 (1047)	3.91 (131)	1.37 (104)	2.82 (688)	1.58 (85)	3.66 (771)	2.93 (107)
vin1	4.85 (1079)	3.76 (134)	1.44 (107)	2.73 (696)	1.71 (85)	3.72 (792)	2.81 (109)
vin2**	5.27 (1079)	4.31 (134)	1.73 (102)	3.45 (696)	2.09 (85)	4.77 (792)	3.91 (109)
vin3**	6.09 (1078)	5.22 (134)	1.71 (101)	3.50 (695)	2.09 (85)	4.79 (792)	3.74 (109)
web	4.00 (1051)	2.85 (133)	1.43 (114)	3.05 (668)	2.03 (84)	3.32 (779)	2.50 (109)
P450						$B < 25.0$ Å^2	
1CPT (22%)§			1.68 (196)				
aba	5.03 (3134)	4.25 (403)	1.89 (311)	4.22 (2099)	3.37 (269)	4.15 (1727)	3.36 (247)
web	7.76 (3135)	7.40 (403)	1.79 (341)	7.00 (2099)	6.59 (269)	6.50 (1727)	5.89 (247)

*For each structure the rmsd (in Å) and the number of atoms used in the superposition (in parentheses) are indicated. For each protein, the variation in the number of atoms used in the least-squares superpositions is due to missing residues (or atoms) in some of the predicted structures.

†α-helical and β-strand residues correspond to those observed in the experimental structure.

‡Atoms with *B*-factors less than a threshold (~1.2 times the mean residue *B*-factor of the experimental structure) were used in the superpositions.

§This is the PDB accession code for the template structure with the highest sequence identity (given in parentheses) to the experimental structure.

**These are fully automated predictions using commercially available Biosym software.

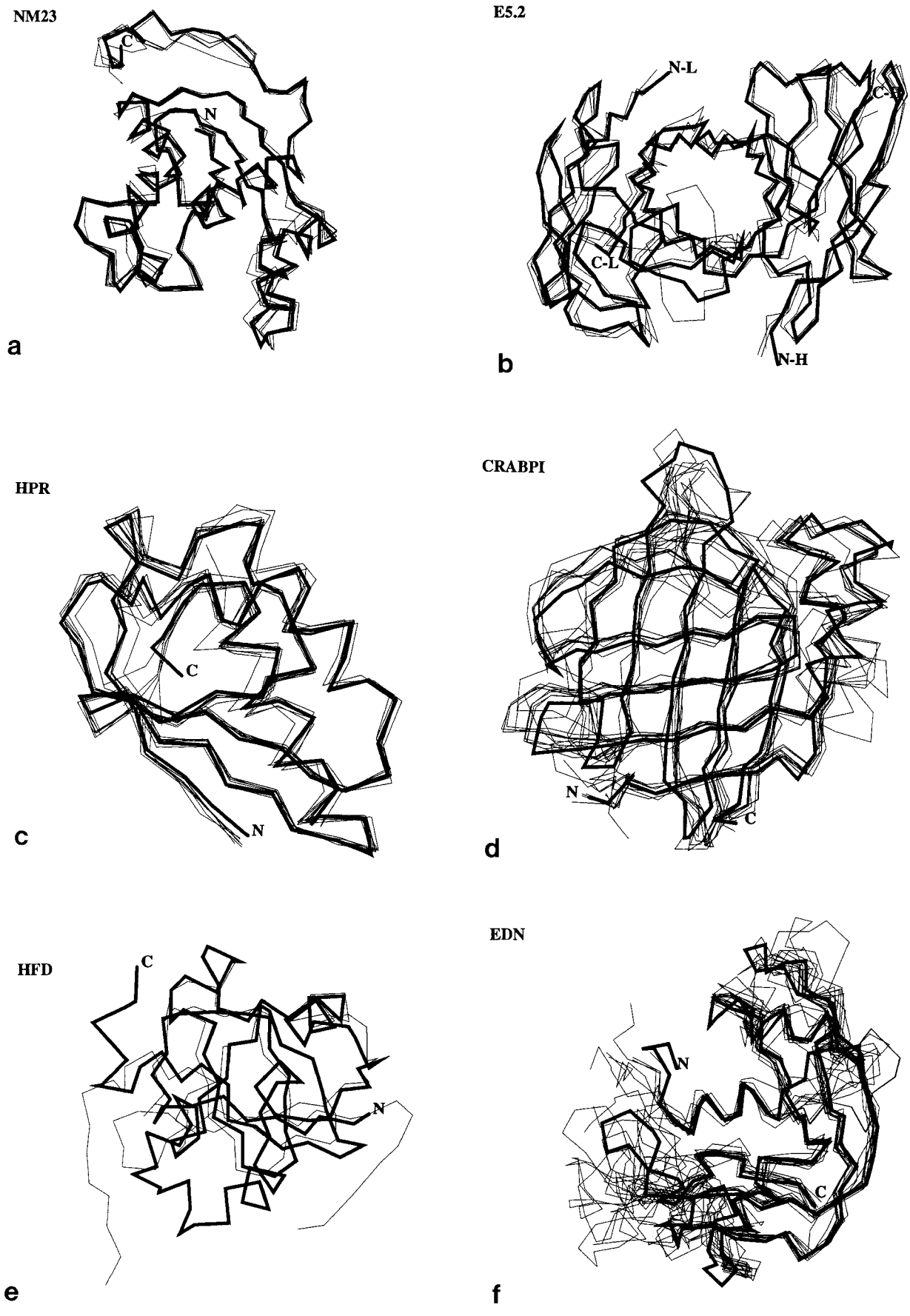

Fig. 1, a-f.

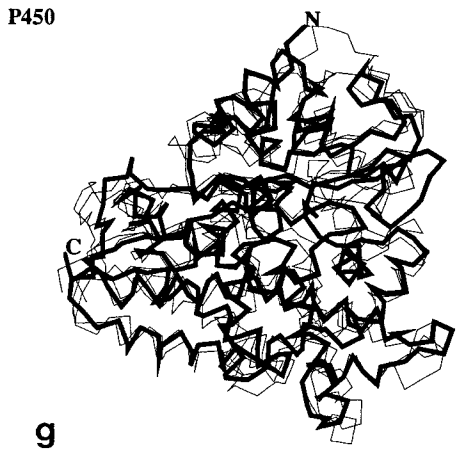

Fig. 1. The seven experimental structures and their corresponding least-squares superimposed comparative molecular models are shown as C^{α} traces. The experimental structures are indicated in bold lines and the comparative molecular models in thin lines. The N- and C-termini of the experimental structures are also indicated. Regions of regular repeating secondary structure are predicted more accurately than the loop regions. The figures are shown in order of decreasing amino acid sequence identity between the experimental structure and its template structure [(**a**) NM23, 77%; (**b**) E5.2, 76%; (**c**) HPR, 42%; (**d**) CRABPI, 41%; (**e**) HFD, 40%; (**f**) EDN, 35%; (**g**) P450, 22%]. In (**a**) the agreement between the experimental NM23 and its predicted structures is excellent. The differences in the conformations of the C-terminal residues result from the use of different template structures. Of the four template structures available for modeling NM23, only 1NDL has a C-terminal conformation equivalent to that found experimentally. In (**e**) and (**g**) the comparative molecular modelers were required to predict structures containing large insertions (> 30 residues). In these cases, the comparative molecular models had erroneous sequence alignments.

also usually the high *B*-factor regions of a protein. In addition, these segments of polypeptide connecting regions of secondary structure are the sites of insertions and deletions in homologous proteins and thus have the lowest values for sequence identity.

Table IV also lists the rmsds for structurally equivalent C^{α} atoms of the experimental structures compared to both the template and modeled structures. Structurally equivalent C^{α} atoms are defined as those C^{α} atoms that are within 3.8 Å in the superposition. For the majority of the predictions of HPR, CRABPI, and EDN, the rmsds on C^{α} atoms for the structurally equivalent residues are of comparable magnitude to the experimental vs. template rmsds on C^{α} atoms. Among the predictions, however, there are some remarkably bad outliers for which the template with no changes would have been a better model of the experimental structure (NM23koe, vih, vri, web1, web2; HPRkoe2; CRABPIvin1, vin2, vin3, web1, web2; EDNbio, vin2, vin3).

HFD (128 residues) and P450 (404 residues) have the largest number of inserted residues relative to the template used, 30 and 101 residues, respectively. The comparative model HFDweb contains a major sequence misalignment that results in the high value of the rmsd (1.69 Å) for the 93 atoms that can be superimposed within 3.8 Å. For P450web the rmsd is 1.79 Å for 341 C^{α} atoms that can be superimposed to within 3.8 Å. Whereas this appears to be better than the P450 least-squares superposition with its template structure (i.e., comparable magnitude for 145 more atom pairs), the P450web prediction has an N-terminal sequence misalignment of 96 residues. The sequence misalignments in P450 and in HFD result in the very high values for the rmsds in the pairwise least-squares superpositions for all C^{α} atoms cases. On the other hand, the very large values for the rmsds on C^{α} atom superpositions for EDN and the predicted structures are due to the large insertions and deletions present in that structure relative to its template molecule.

Peptide bonds in proteins, for the most part, are close to being planar. Thus, departures from peptide bond planarity (Table III and Table V) can be used to evaluate the quality of the predictions. In well-refined protein structures few peptide bonds have ω angles that deviate by more than $\pm$ 6° from 180° (or from 0° for *cis*-peptides). A surprisingly large number of the predictions had a large percentage of the peptide bonds that deviated by more than $\pm$ 15° from the values in the experimental structures (Table V). Those models with the largest percentages of nonplanar peptide bonds (here we mean poor agreement with the experimental structures) also had the largest absolute differences. In these, the trigonal character of the carbonyl-carbon atom was totally lost as was the hydrogen-bonding normally associated with peptide bonds in secondary structures. These large departures from planarity are far more frequent in many of the structures than one would expect from poor geometry at a few splice points. For laboratories with consistently large departures from peptide bond planarity (aba, mou, vin, web) the energy minimization protocols should be examined.

Table V also lists the percentage of residues in the predictions that deviate in χ_1 and χ_2 from the experimental structures by more than $\pm$ 30°. In general, χ_2 is more poorly predicted than χ_1, but even the best predictions (NM23sal) still have major differences in the χ_1 and χ_2 torsion angles. Some of the predictions (EDNbio, saq2, vin2, vin3, and web) had only 33 to 42% of the side chains with correctly assigned values of χ_2. Clearly side chain torsion angles are poorly predicted. Unless the side chains are from the set of conserved residues in the template structure, and associated with the active site or in the hydrophobic interior of the molecule, the predicted structures score poorly on the values of χ_1 and χ_2.

DISCUSSION

The prediction of a target protein's conformation based on a related template structure can be highly accurate for proteins sharing a moderate to high degree of amino acid sequence identity and not having

TABLE V. Comparison of the Dihedral Angles in the Experimental Structure and the Comparative Molecular Models

NM23

	aba	koe	sal	vih	vri	web1	web2
Percent $\Delta\omega > \pm 15°$	6	1	0	3	3	32	25
Max. $\Delta\omega$ (°)	54.2	16.1	11.7	25.0	139.7	113.1	58.5
Percent $\Delta\chi_1 > \pm 30°$	29	35	25	47	40	44	47
Percent $\Delta\chi_2 > \pm 30°$	35	45	34	46	43	56	53

E5.2

	aba	bol*	oxm*
Percent $\Delta\omega > \pm 15°$	27	5	5
Max. $\Delta\omega$ (°)	76.0	177.6	178.3
Percent $\Delta\chi_1 > \pm 30°$	50	36	43
Percent $\Delta\chi_2 > \pm 30°$	50	56	52

HPR

	aba	bio	koe1	koe2	mou	tri	vih	vri	web
Percent $\Delta\omega > \pm 15°$	36	5	0	0	19	0	0	2	23
Max. $\Delta\omega$ (°)	62.4	18.4	14.1	9.5	29.0	11.7	14.1	21.2	32.3
Percent $\Delta\chi_1 > \pm 30°$	32	42	41	41	41	39	51	36	45
Percent $\Delta\chi_2 > \pm 30°$	52	46	50	51	59	54	62	61	46

CRABPI

	aba	mou1	mou2	sal	vin1	vin2†	vin3†	vri	web1	web2
Percent $\Delta\omega > \pm 15°$	15	7	16	1	16	13	17	1	52	60
Max. $\Delta\omega$ (°)	58.8	98.9	57.5	15.8	40.2	27.4	40.1	16.2	90.3	74.5
Percent $\Delta\chi_1 > \pm 30°$	49	47	47	39	53	60	53	39	46	49
Percent $\Delta\chi_2 > \pm 30°$	49	39	40	44	59	61	63	37	55	54

HFD

	web
Percent $\Delta\omega > \pm 15°$	45
Max. $\Delta\omega$ (°)	139.6
Percent $\Delta\chi_1 > \pm 30°$	49
Percent $\Delta\chi_2 > \pm 30°$	68

EDN

	bio	koe	mou†	sal1	sal2	saq1‡	saq2	vin1‡	vin2†	vin3†	web
Percent $\Delta\omega > \pm 15°$	32	5	32	0	0	8	11	18	21	18	52
Max. $\Delta\omega$ (°)	41.2	20.7	159.6	11.4	13.1	175.7	76.7	179.5	34.4	27.0	136.6
Percent $\Delta\chi_1 > \pm 30°$	53	51	44	35	35	47	42	41	51	58	49
Percent $\Delta\chi_2 > \pm 30°$	58	45	53	49	49	54	67	52	61	64	67

P450

	aba	web
Percent $\Delta\omega > \pm 15°$	10	45
Max. $\Delta\omega$ (°)	130.2	145.9
Percent $\Delta\chi_1 > \pm 30°$	59	53
Percent $\Delta\chi_2 > \pm 30°$	58	56

*The large deviations in ω for these predictions result from the comparison of a *cis*-peptide (experimental structure) and modeled *trans*-peptides.

†These are fully automated predictions using commercially available Biosym software.

‡The large deviations in ω for these predictions results from the comparison of a *trans*-peptide (experimental structure) and modeled *cis*-peptides.

large insertions or deletions relative to the template. NM23 is 77% identical to the template, 1NDL, and both have 148 residues per subunit. The rmsd on C^{α} atoms of 0.53 Å between subunit N of NM23 and NM23sal is comparable in size to the differences observed in pairwise comparisons among the six identical subunits of NM23 (rmsds range from 0.30 to 0.57 Å). It is also well within the range of rmsds observed when comparing RNase A molecules whose structures were determined in different space groups (0.16–0.79 Å).[52] The comparative models of HPR (40% amino acid sequence identity and one additional C-terminal residue) and E5.2 (when considering the subunits independently) are similar to the NM23 models in quality. Model building of proteins from template protein structures having lower sequence identity and/or containing insertions or deletions is more difficult and the results are significantly less accurate. This is evident from the rmsds in Table IV and can be seen qualitatively in Figure

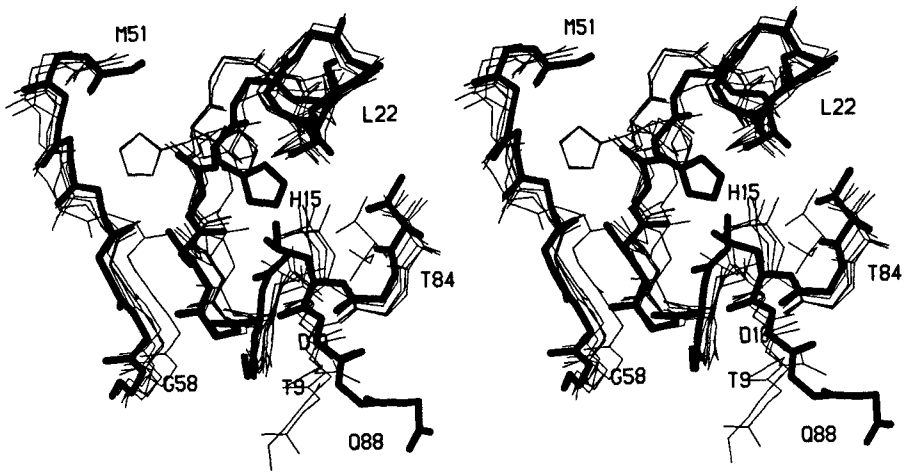

Fig. 2. The HPR active site is shown for the experimental structure (bold lines) and for five of the nine comparative molecular models (thin lines). The position of the loop bearing His-15 in the experimental structure is displaced relative to the loop position in the template structure (not shown). The comparative molecular models follow the path in the template. The imidazole ring of His-15 in the experimental structure is ~2.5 Å from the position of the equivalent ring in the closest of the models. The differences are due to the closer approach of the N-terminal β-strand (residue 1 to 8) and helix α1 in HPR than in the template structure (2HPR). Another potential explanation for the observed difference is the hydrogen-bonding interaction between His-15 and Asp-10. In the template structure, Asp-10 is an alanine and does not contact His-15 directly. Note that two of the predicted structures have χ_1 conformations of His-15 that do not correspond to either the experimental or template structures.

1e–g. In these cases, the template structure does not accurately predict the conformation of parts of the experimental structure. This is particularly true when there are substantial insertions or deletions between the target and template structures. Often these differences are larger than the radius of convergence of energy minimization techniques. Even worse, the insertions and deletions between the target and template structures can lead to errors in sequence alignment and therefore unavoidable errors in the molecular model.

From the results gathered here, it appears that present modeling procedures followed by energy minimization are limited in their ability to move the predicted structure toward the observed crystallographic structure. For example, the loop in HPR connecting the N-terminal β-strand (residues 1–8) to helix α1 contains the catalytic residue His-15. Nine of the 14 residues between Thr-8 and Leu-22 are identical in HPR and its template structure (2HPR). Despite this high degree of amino acid sequence identity and the lack of inserted or deleted residues, His-15 and supporting parts of this loop in HPR are shifted relative to the same loop in the template structure (Fig. 2). The 10 predicted structures faithfully follow the template structure thereby failing to predict the correct position of His-15 in HPR. It is not entirely obvious why HPR differs in this region from its template structure, 2HPR. In HPR, the N-terminal β-strand and helix α1 approach more closely than observed in the template structure, as the side chains that pack between the secondary structural units are less bulky than those in 2HPR. The carboxylate of Asp-10 participates in an Asx turn by accepting a hydrogen bond from the main-chain NH of Thr-12. Also, in HPR, His-15 donates a hydrogen bond to the carboxylate of Asp-10 facilitating their close approach. In the parent structure, 2HPR, the equivalent residue to Asp-10 is an alanine that does not interact with the His-15 side chain. The imidazole rings are separated by ~2.5 Å in the least-squares superposition of HPR on its parent structure, 2HPR.

In the structure-based sequence alignment of EDN and its template structure (RNase A), the connecting polypeptide segment between helix α1 and helix α2 is six residues shorter in EDN than in RNase A. The conformational differences in this region are large (Fig. 3). In RNase A, helix α2 begins one turn earlier than the equivalent helix in EDN. The large side chains of Tyr-23, Met-29, and Arg-33 are on the same side of the α2 helix where they pack against the loop connecting helix α1 and helix α2. In EDN, the equivalent residues are Gln-22, Ala-28, and Val-31. Gln-22 is not part of the α2 helix in EDN and its ψ angle differs by ~120° when compared to the equivalent residue, Tyr-23, in RNase A. The side chain of Glu-22 faces the polar solvent as a result of this different main chain conformation. Ala-28 and Val-31 in EDN are significantly smaller than their counterparts in RNase A. As a consequence, the EDN loop occupies the space that is filled by the large side chains of Tyr-23, Met-29, and Arg-33 in RNase A.

In Figure 4, the same loop and associated helices from the EDNsal1 prediction are shown. In this pre-

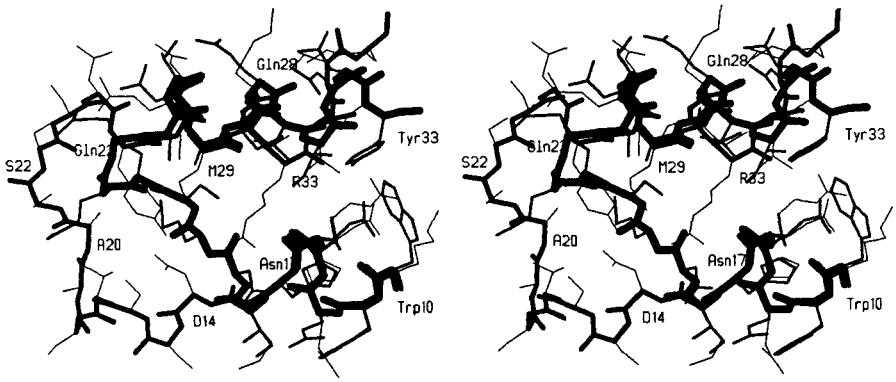

Fig. 3. A stereodiagram representing the loop between the helices α1 and α2 in EDN (thick lines) and RNase A (thin lines). EDN has a 6 residue deletion in this loop relative to RNase A. Helix α2 in EDN is one turn shorter than in RNase A. The bulky side chains of Tyr-25, Met-29, and Arg-33 on helix α2 of RNase A pack against this loop in the template structure. In EDN, the equivalent residues are the smaller Gln-22, Ala-26, and Val-29. As a result, the loop in EDN adopts a conformation that fills the space between helices α1 and α2.

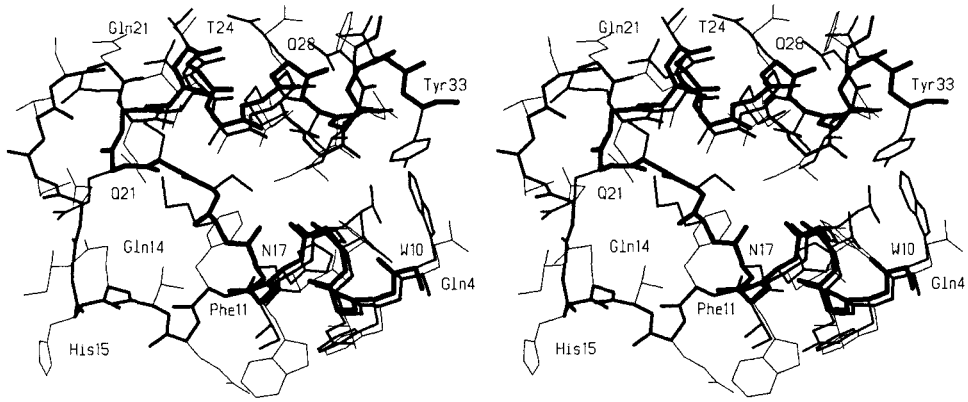

Fig. 4. A stereodiagram representing the N-terminal residues in the superposition of the experimental EDN and the comparative molecular model EDNsal1. This figure shows the result of an incorrect sequence alignment of EDN and RNase A on the comparative model. In this case, the 6 residue deletion in EDN was incorrectly aligned and the EDNsal1 model follows the RNase A conformation for this loop. A major consequence of the misalignment is a change in residue type for the predicted active site residues. The Gln-14 and His-15 residues of EDN are replaced by Ala-8 and Gln-9. This change cannot be justified in terms of the accepted RNase A mechanism of hydrolysis.

diction, the six residue deletion between the two helices α1 and α2 has been placed at the N-terminus so that residues 1–16 in EDNsal1 are misaligned. In this case the misalignment changes the nature of two active site residues from Gln-14–His-15 (EDN) to Ala-8–Gln-9 (EDNsal1). The active site His-15 in EDNsal1 is ~22 Å from its true position in the EDN active site. Such errors in alignment could be avoided by paying close attention to the nature of catalytically important residues in these enzymes. Typically, sequence alignment errors give rise to the largest rmsds observed in comparisons of the model and target structures. Thus, the very large rmsds listed in Table IV are due to sequence alignment errors.

The comparative molecular models have failed completely to predict the correct conformations of the large insertions in EDN, HFD, and P450. In EDN, there is a nine residue insertion between strands β5 and β6 (the inserted residues are from Asp-115 to Tyr-123). The conformation that this loop adopts is shown in Figure 5a. Residues Asp-115 to Asp-119 are in a single turn of helix that is initiated by the Asp-115 side chain forming a hydrogen bond to the NH of Arg-118. The Asp-119 side chain is directed toward the N-terminus of helix α1 where its negative carboxylate stabilizes the positive pole of the α1 helix dipole. The guanidinium group of Arg-114 donates hydrogen bonds to the Asp-119 O (3.0 Å) and Gln-116 O (2.9 Å) thereby satisfying the negative pole of the single helical turn and breaking this short helix. Pro-120 to Tyr-123 are solvent accessible and interact with ordered solvent molecules. The nine residue loop has well-defined associated elec-

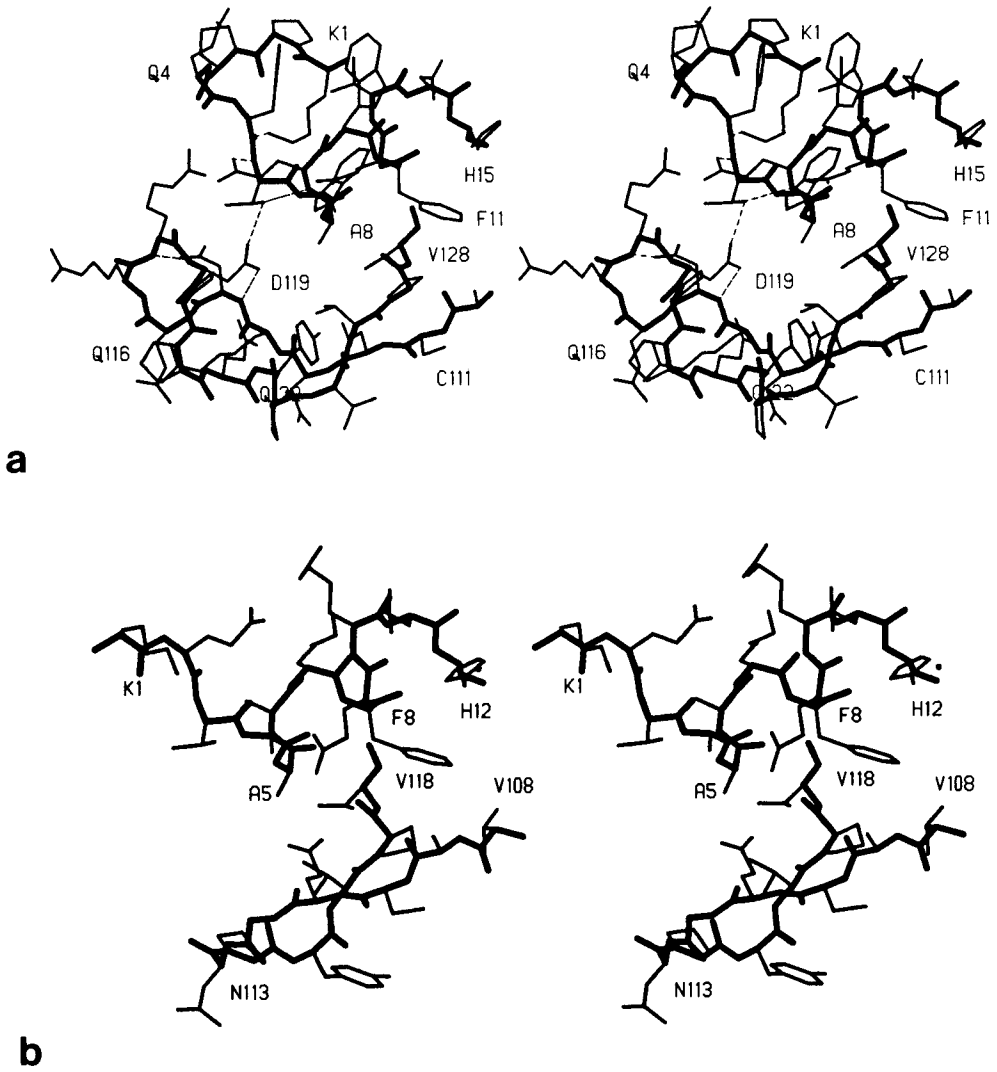

Fig. 5. (**a**) A stereodiagram showing the 9 residue insertion (Asp-115–Tyr-123) between strands β5 and β6 of EDN relative to RNase A. The inserted residues form a single turn of helix and help to stabilize the conformation of the N-terminal helix. Asp-115 initiates the single turn of helix and the carboxylate of Asp-119 opposes the positive pole of the N-terminal helix dipole. The equivalent region of RNase A is shown in (**b**). In both instances the heavy lines correspond to the main chain peptide bonds and the thin lines represent the side chains. Intramolecular hydrogen bonds in (**a**) are indicated as dashed lines.

tron density and is not involved in any significant direct crystal contacts. The helical conformation of Asp-115–Asp-119 and the Asp-119 interaction with the N-terminal helix were not predicted in any of the 11 submitted EDN predicted structures. The conformation of the first four residues of helix α1 was not correctly predicted (Fig. 1f). Instead of a continuous helix from Lys-1 to Ile-16 the first five residues Lys-1 to Phe-5 fold back on the helix and it is initiated at Thr-6. Figure 5a shows that these two regions (the N-terminus and the nine residue insertion) pack against one another. This synergy makes it exceedingly difficult to predict this region. The equivalent region of RNase A is shown in Figure 5b.

One of the comparative modeling groups submitted their two interactively built models (CRABPIvin1, EDNvin1) along with "fully automated" predictions of these structures using the commercially available Consensus software from Biosym (CRABPIvin2, vin3, EDNvin2, vin3). A comparison of these results shows that the fully automatically constructed models have larger rmsds for the C^{α} and all atom superpositions than do the interactively built models. For EDN the magnitude of the differences is larger than for CRABPI. It would appear that more careful attention needs to be paid when using any fully automated off-the-shelf software.

It should be noted that the rmsds for the all atom and C^{α} atom superpositions are only slightly affected by improper stereochemistry. Predictions having high percentages of nonplanar ω angles all have rmsds that are changed only slightly after reg-

ularization of the stereochemistry. Similarly, structures with a significant number of D-amino acids have rmsds that differ by only a few 10ths of an Å after the conversion to L-stereochemistry followed by regularization. This does not suggest nonplanar ω angles and D-amino acids have little effect on the quality of the resulting model. It points out a major shortcoming of the blind use of the rmsd as a criterion for evaluating comparative molecular models.

Each of the EDN comparative molecular models was tested as rotation function search models to see if they were able to provide a molecular replacement solution for EDN. Using the AMORE program suite,[53] EDNbio (peak 13), EDNsaq1 (peak 11), and EDNsaq2 (peak 14) are the only models that give the correct solution as one of the top 25 rotation function peaks. The EDN structure was solved using a truncated version of RNase A as the whole RNase A molecule did not produce a molecular replacement solution among the top 50 peaks (Mosimann et al., unpublished results).

If comparative molecular models are to be used to examine detailed differences in substrate, ligand, or receptor association, future comparative molecular modelers will have to improve the stereochemical quality of their models. This can easily be achieved by the standardized usage of some form of structure verification program (e.g., PROCHECK[45]). Experimentally determined protein structures are routinely assessed with structure verification programs both during and after the structure determination in order to identify potential problem regions in the structure. Structure verification programs can aid the comparative molecular model builder by evaluating the stereochemical quality of proposed conformations in their models. In the submitted comparative molecular models, the "splice" points or segments containing large insertions or deletions are the regions that most often have residues with unlikely stereochemical parameters.

The present exercise has brought to light a number of problems that still exist in comparative molecular modeling. The availability of the experimental structures has allowed us to carry out critical assessment of the predicted structures. When comparative molecular model coordinates are submitted to databases or when papers describing the results of the modeling are submitted to journals the experimental structures will not be available. We feel that any coordinates of comparative molecular models submitted for possible publication should be accompanied by the output from a structure verification program that will allow for an objective assessment of the stereochemical quality of the model. The model should have a distribution of ϕ,ψ angles that is similar to those expected for experimental structures. The model should have few, if any, unsatisfied hydrogen bond donors or acceptors in solvent inaccessible regions of the molecule. Solvent accessible surfaces should be primarily populated by hydrophilic groups or atoms. In a simplistic fashion this is saying that hydrophobic groups in the model should be on the inside and hydrophilic groups should be solvent accessible. Nonbonded contacts that are less than acceptable limits observed in real protein structures should be eliminated (i.e., there should be no intramolecular contacts less than 2.5 Å in the protein model). The distribution of side chain conformations, χ_n, should fall in the preferred regions that have been tabulated for the individual residue types from analyses of databases.[47] We have found that most of these requirements can be met by molecular modelers with the use of the PROCHECK program.

ACKNOWLEDGMENTS

We thank the Medical Research Council of Canada for continued funding to the Group in Protein Structure and Function that has allowed us to do these evaluations. MNGJ thanks the Lawrence Livemore National Laboratory for a travel grant to attend the conference at Asilomar. S.M. acknowledges the Alberta Heritage Foundation for Medical Research for the generous award of a studentship.

REFERENCES

1. Greer, J. Comparative model-building of the mammalian serine proteases. J. Mol. Biol. 153:1027–1042, 1981.
2. Greer, J. Comparative modeling methods: Application to the family of the mammalian serine proteases. Proteins 7:317–334, 1990.
3. Murphy, M.E.P., Moult, J., Bleackley, R.C., Weissman, I.L., James, M.N.G. Comparative molecular model building of two serine proteinases from cytotoxic T lymphocytes. Proteins 4:190–204, 1988.
4. Carson, M., Bugg, C.E., Delucas, L., Narayana, S. Comparison of homology models with the experimental structure of a novel serine protease. Acta Crystallogr. D50:889–899, 1994.
5. Li, Z., Chen, X., Davidson, E., Zwang, O., Mendis, C., Ring, C., Roush, W.R., Fegley, G., Li, R., Rosenthal, R.J., Lee, G.K., Kenyon, G.L., Kuntz, I.D., Cohen, F.E. Anti-malarial drug development using models of enzyme structure. Chem. Biol. 1:31–37, 1994.
6. Odake, S., Kam, C.M., Narasimhan, L., Poe, M., Blake, J.T., Krahenbuhl, O., Tschopp, J., Powers, J.C. Human and murine cytotoxic T lymphocyte serine proteases: Subsite mapping with peptide thioester substrates and inhibition of enzyme activity and cytolysis by isocoumarins. Biochemistry 30:2217–2227, 1991.
7. Caputo, A., James, M.N.G., Powers, J.C., Hudig, D., Bleackley, R.C. Conversion of the substrate specificity of mouse proteinase granzyme B. Nature Struct. Biol. 1:364–367, 1994.
8. Acharya, K.R., Stuart, D.I., Phillips, D.C., Scheraga, H.A. A critical evaluation of the predicted and X-ray structures of α-lactalbumin. J. Prot. Chem. 9:549–563, 1990.
9. Browne, W.J., North, A.C., Phillips, D.C. A possible three-dimensional structure of bovine α-lactalbumin based upon that of hen's egg-white lysozyme. J. Mol. Biol. 42(1):65–86, 1969.
10. Warme, P.K., Momany, F.A., Rumbell, S.V., Tuttle, R.W., Scheraga, H.A. Computation of structures of homologous proteins: α-lactalbumin from lysozyme. Biochemistry 13: 768–782, 1974.
11. Blake, C.C.F., Koenig, D.F., Mair, G.A., North, A.C.T., Phillips, D.C., Sarma, V. Structure of hen egg-white lysozyme, a three-dimensional Fourier synthesis at 2 Å resolution. Nature (London) 206:757–761, 1965.

12. Shotton, D.M., Watson, H.C. Three-dimensional structure of tosyl-elastase. Nature (London) 225:811–816, 1970.
13. Matthews, B.W., Sigler, P.B., Henderson, R., Blow, D.M. Three-dimensional structure of tosyl-α-chymotrypsin. Nature (London) 214:652–656, 1967.
14. McLachlan, A.D., Shotton, D.M. Structural similarities between α-lytic protease of Myxobacter 495 and elastase. Nature (London) 229:202–205, 1971.
15. Delbaere, L.T.J., Brayer, G.D., James, M.N.G. Comparison of the predicted model of α-lytic protease with the X-ray structure. Nature (London) 279:165–168, 1979.
16. Brayer, G.D., Delbaere, L.T.J., James, M.N.G. Molecular structure of the α-lytic protease from *Myxobacter* 495 at 2.8 Å resolution. J. Mol. Biol. 131:743–775, 1979.
17. James, M.N.G., Delbaere, L.T.J., Brayer, G.D. Amino acid sequence alignment of bacterial and mammalian pancreatic serine proteases based on topological equivalences. Can. J. Biochem. 56:396–402, 1978.
18. Read, R.J., Brayer, G.D., Jurasek, L., James, M.N.G. Critical evaluation of comparative model-building of *Streptomyces griseus* trypsin. Biochemistry 23:6570–6575, 1984.
19. Bing, D.H., Laura, R., Robison, D.J., Furie, B., Furie, B.C., Feldmann, R.J. A computer-generated three-dimensional model of the B chain of bovine α-thrombin. Ann. N.Y. Acad. Sci. 370:496–510, 1981.
20. Bode, W., Mayr, I., Baumann, U., Huber, R., Stone, S.A., Hofsteenge, J. The refined 1.9 Å crystal structure of human α-thrombin:interaction with D-Phe-Pro-Arg chloromethylketone and significance of the Tyr-Pro-Pro-Trp insertion segment. EMBO J. 8:3467–3475, 1989.
21. Rydel, T.J., Tulinsky, A., Bode, W., Huber, R. Refined structure of the hirudin-thrombin complex. J. Mol. Biol. 221:583–601, 1991.
22. Zdanov, A., Wu, S., DiMaio, J., Konishi, Y., Li, Y., Wu, X., Edwards, B.F., Martin, P.D., Cygler, M. Crystal structure of the complex of human α-thrombin and nonhydrolyzable bifunctional inhibitors, hirutonin-2 and hirutonin-6. Proteins 17:252–265, 1993.
23. Chothia, C., Lesk, A.M. Evolution of proteins formed by β-sheets. I. Plastocyanin and azurin. J. Mol. Biol. 160:309–323, 1982.
24. Lesk, A.M., Chothia, C. Evolution of proteins formed by β-sheets. II. The core of the immunoglobulin domains. J. Mol. Biol. 160:325–342, 1982.
25. Blundell, T., Sibanda, B.L., Pearl, L. 3-D structure, specificity and catalytic mechanism of renin. Nature (London) 304:273–275, 1983.
26. Sielecki, A.R., Hayakawa, K., Fujinaga, M., Murphy, M.E.P., Fraser, M., Muir, A.K., Carilli, C.T., Lewicki, J.A., Baxter, J.D., James, M.N.G. Molecular structure of a target for cardiovascular-active drugs: Recombinant human renin at 2.5 Å resolution. Science 243:1346–1351, 1989.
27. Kleywegt, G.J., Bergfors, T., Senn, H., Le Motte, P., Gsell, B., Shudo, K., Jones, T.A. Crystal structure of cellular retinoic acid binding proteins I and II in complex with all-trans-retinoic acid and a synthetic retinoid. Structure 2:1241–1258, 1994.
28. Cupp-Vickery, J.R., Poulos, T.L. Structure of cytochrome P450eryF involved in erythromycin biosynthesis. Nature Struct. Biol. 2:144–153, 1995.
29. Bernstein, F.C., Koetzle, T.F, Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasmui, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
30. Abagyan, R., Totrov, M., Kuznetsov, D. ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. J. Comp. Chem. 15:488–506, 1994.
31. Bolger, M.B., Sherman, M.A. Computer modeling of combining site structure of anti-hapten monoclonal antibodies. Methods Enzymol. 203:21–45, 1991.
32. Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta Jr., S., Weiner, P. A new force field for molecular mechanical simulation of nucleic acids and proteins. J. Am. Chem. Soc. 106:765–784, 1984.
33. BIOSYM Software. Biosym Technologies, Inc., INSIGHTII Software Suite, 9685 Scranton Road, San Diego, CA. 92121–2777.
34. Koehl, P., Delarue, M. A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. Nature Struct. Biol. 2:163–170, 1995.
35. Samudrala, R., Pedersen, J.T., Zhou, H., Luo, R., Fidelis, K., Moult, J. Confronting the problem of correlated structural change in comparative modeling proteins. Proteins, 23:000–000, 1995.
36. AbM, Oxford Molecular, Suite 300, 700 East El Camino Road, Mountain View, CA. 94040.
37. Sali, A., Blundell, T.L. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234:779–815, 1993.
38. QUANTA Software Suite. Molecular Simulations, Inc., 16 New England Executive Park, Burlington, MA. 01803.
39. Peitsch, M.C., Jongeneel, C.V. A 3-dimensional model for the CD40 ligand reveals a close similarity to the tumor necrosis factors. Int. Immunol. 5:233–238, 1993.
40. TRIPOS Software Suite, TRIPOS, Inc., 1699 South Hanley Road, Suite 303, St. Louis, Missouri 63144.
41. Vihinen, M., Euranto, A., Luostarinen, P., Nevalainen, O. MULTICOMP: A program package for multiple sequence comparison. Comput. Appl. Biosci. 8:35–38, 1992.
42. Vinals, C., De Bolle, X., Depiereux, E., Feytmans, E. Knowledge-based modeling of the D-lactate dehydrogenase three-dimensional structure. Proteins 21:307–318, 1995.
43. Vriend, G. WHAT IF: A molecular modeling and drug design program. J. Mol. Graphics 8:52–56, 1990.
44. Harrison, R.W., Weber, I.T. Molecular dynamics simulation of HIV-1 protease with a peptide substrate. Protein Eng. 7:1353–1363, 1994.
45. Laskowski, R.A., MacArthur, M.W., Moss, D.S., Thornton, J.M. PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst. 26:283–291, 1993.
46. Jones, T.A., Zou, J.Y., Cowan, S.W., Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Cryst. A47:110–119, 1991.
47. Dunbrack, R., Karplus, M. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nature Struct. Biol. 1:334–340, 1994.
48. Aho, A.V., Kernighan, B.W., Weinberger, P.J. "The AWK Programming Language." Reading, MA: Addison-Wesley, 1988.
49. Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Cryst. A34:827–828, 1978.
50. McLachlan, A.D. Gene duplication in the structural evolution of chymotrypsin. J. Mol. Biol. 128:49–79, 1979.
51. Kabsch, W., Sander, C. Dictionary of protein secondary structure:pattern recognition of hydrogen bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
52. Zegers, I., Maes, D., Thi, M.H.D., Wyns, L., Poortmans, F., Palmer, R. The structure of RNase A complexed with 3′-CMP and d(CpA):Active site conformation and conserved water molecules. Protein Sci. 3:2322–2339, 1994.
53. Navaza, J. AMoRe: An automated package for molecular replacement. Acta Cryst. A50:157–163, 1994.