PROTEINS: Structure, Function, and Genetics 5:170–182 (1989)

Rebuilding Flavodoxin From Cα Coordinates: A Test Study

Lorne S. Reid and Janet M. Thornton
Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, London, England WC1E 7HX

ABSTRACT **The tertiary structure of flavodoxin has been model built from only the X-ray crystallographic α-carbon coordinates. Main-chain atoms were generated from a dictionary of backbone structures. Side-chain conformations were initially set according to observed statistical distributions, clashes were resolved with reference to other knowledge-based parameters, and finally, energy minimization was applied. The RMSD of the model was 1.7 Å across all atoms to the native structure. Regular secondary structural elements were modeled more accurately than other regions. About 40% of the χ_1 torsional angles were modeled correctly. Packing of side chains in the core was energetically stable but diverged significantly from the native structure in some regions.**

The modeling of protein structures is increasing in popularity but relatively few checks have been applied to determine the accuracy of the approach. In this work a variety of parameters have been examined. It was found that close contacts, and hydrogen-bonding patterns could identify poorly packed residues. These tests, however, did not indicate which residues had a conformation different from the native structure or how to move such residues to bring them into agreement. To assist in the modeling of interacting side chains a database of known interactions has been prepared.

Key words: modeling, flavodoxin, structure prediction, side chains, database, structure analysis, protein

INTRODUCTION

Protein modeling is being used increasingly for structure prediction.[1–3] Most often this modeling is performed on the basis of homology to a protein of known structure. Difficulties in the method increase as the percentage homology between the two proteins decreases.[4,5] Most studies have concentrated on backbone conformation, with root mean square deviations (RMSD) expressed for Cα atoms only.[6] However, most changes in backbone conformation and intermolecular recognition are mediated by side-chain interactions. This paper explores our ability to predict side-chain, and to a lesser extent, main-chain conformation, given only α-carbon coordinates. It is an idealized "experiment" because in the modeling of an homologous protein the backbone conformation will also be modified by changes in sequence. This work was stimulated by an attempt to model build an homologous protein on the basis of Cα coordinates, where all the atomic coordinates were unavailable. Flavodoxin was used as a test structure because (1) it is known to a resolution of 1.9 Å with a R factor = 21.4%,[7] (2) it contains a mixture of α-helix, β-sheet, turns, coil conformations, and an active site (FMN binding domain), (3) all amino acid types, except histidine, are present, and (4) it is a moderate size of 138 residues.

In proteins of high homology the side chain χ_1 angles are mostly retained. As recently reported, for proteins with greater than 40% sequence homology the side-chain γ atoms of homologous residues show 80–97% retention of configuration; the δ atoms have a 75% or more transferability of position.[8] Surface accessibility is an important factor when modeling homologous changes; nonhomologous, solvent accessible pairs show random orientation. However, these data are relevant only in those situations where the backbone structural integrity is maintained during substitutions. Side-chain modeling must be approached ab initio where homology is low or where an altered backbone conformation is proposed (e.g., because of insertions or deletions). Also, this work has a broader relevance because the Cα trace is determined first from the electron density map during structure analysis; prediction of side-chain conformations could aid in map interpretation. The procedure is also relevant to ab initio prediction of structure where a novel sequence may be fitted to a coordinate template (e.g., a 4-helical bundle template) and side-chain conformations and interactions will determine whether the structure is "acceptable."

Many protein structure models have appeared in

Received June 23, 1988; accepted February 1, 1989

Address reprint requests to Lorne S. Reid, Allelix Biopharmaceuticals, 6850 Goreway Dr., Mississauga, Ontario, Canada L4V 1P1.

recent literature. One problem encountered in such work until recently has been the lack of independent methods to determine the validity of the model.[9] In those few cases where the X-ray structure has subsequently been determined the model was found to have essentially the correct backbone fold. However, the detail associated with the placement of side-chain torsional angles was often in error.[6]

This work describes a number of parameters that can be checked and these may point to regions of the structure that deviate from the statistical average. Such deviation does not necessarily imply though that the structure is, in fact, incorrect. Further, these tests, on their own, give no clue as to how to improve the configuration of a particular side chain.

To assist in the modeling of side-chain conformations a database of the three-dimensional interactions of side chains as found in well-refined protein structures has been established. During the modeling of a particular pair of interacting side chains the database is searched for all examples where the backbone is in the geometry required for the model. The results are displayed graphically and the side chains can then be built into the occupied volume.

METHOD

During the model building procedure outlined below the Cα atomic coordinates were kept fixed to their original X-ray values except for the final cycles of the energy minimization step. The FMN moiety was not modeled as this would have required additional information about the location of the active site, something which may not be available in a real modeling situation.

Model Building Procedure

Step 1: The Cα atomic coordinates and residue number (to identify chain direction) were extracted from the Brookhaven Protein Databank entry (3FXN) for *Clostridium* flavodoxin.[10]

A polyalanine backbone was generated with the backbone dictionary option (DGNL) of the FRODO (version Tom 4.0 based on Rice 6.0) graphics package[11] operating on an Evans and Sutherland PS330 unit. With this method short segments, 4–7 residues, of backbone were extracted from a dictionary of 31 protein crystal structures. The dictionary did include coordinates for flavodoxin but these were excluded from consideration. The backbone which fit to the Cα coordinates positions with the lowest RMSD (generally found to be $\leq$ 0.1 Å) was taken as the starting backbone structure for that region. Segments of backbone were not overlapped unless the RMSD was $\geq$ 0.3 Å or there appeared to be a problem with the fitting (e.g., incorrect hydrogen bonding patterns in α-helices or β-sheets).

Step 2: The REPLACE function of FRODO was used to generate side-chain coordinates of idealized geometry in a fully extended conformation.

Step 3: All bond lengths and backbone angles except ϕ and ψ were regularized to ideal values with the REFI option of FRODO.

Step 4: All nonglycine residues were confined to the α, β, and α_L regions of the Ramachandran plot. Glycine residues were allowed to plot into any region. The ϕ and ψ angles of irregular residues were adjusted by refitting the backbone, at these sites, as described in Step 1.

Step 5: Using a restricted subset of the main-chain plus Cβ atoms all interatomic distances of $\leq$ 2.5 Å were identified. These clashes were resolved by choosing a different backbone conformation from the dictionary as in Step 1.

Step 6: The side-chain torsional angles were manually adjusted, by the TOR option of FRODO, to reflect the primary values observed in a statistical analysis of side-chain torsional angles as a function of secondary structure.[12] No attempt was made to resolve clashes between side chains until the preferential χ angles had been set for all residues.

Step 7: New interatomic van der Waals clashes that originated from side-chain atoms were then identified. These clashes were resolved by adjusting the torsional angles of the relevant side chains to reflect less populated zones of the χ_1, χ_2 distribution profiles. In the case of interacting side chains it was often found that one side chain had a single, preferentially populated zone while its partner had two or three favorable substates.

Step 8: In keeping with the knowledge-based approach to model building the side-chain conformations were checked by several criteria. In particular, the angle of approach of aromatic groups,[13–15] and hydrogen-bonding patterns in turns[2] were checked and conformations adjusted accordingly.

As a tool to be used in the modeling procedure a database was constructed of all pairwise side-chain interactions in 159 proteins in the Brookhaven Protein Databank (release October 1987) that had a resolution of 2.4 Å or better. Homologous structures were incorporated in this database to accommodate sequence variations and differences in side-chain interactions. Two residues were classified as interacting if any two atoms (one from each residue) were within 4.0 Å. An index to these interactions was compiled and a Fortran program, PAIRS, written to access the Brookhaven coordinate database. As input, the program requires the main-chain coordinates of two residues that are being modeled. The database is searched for all examples of equivalent pairs of amino acids whose main chains were superimposable on those of the model. Results are viewed graphically on an Evans and Sutherland PS330 unit running an appropriate graphics package such as FRODO or HYDRA.

Step 9: A Cartesian-space, conjugant gradient energy minimizer (program EMPMDS) was then applied to the model structure.[16] An all atom param-

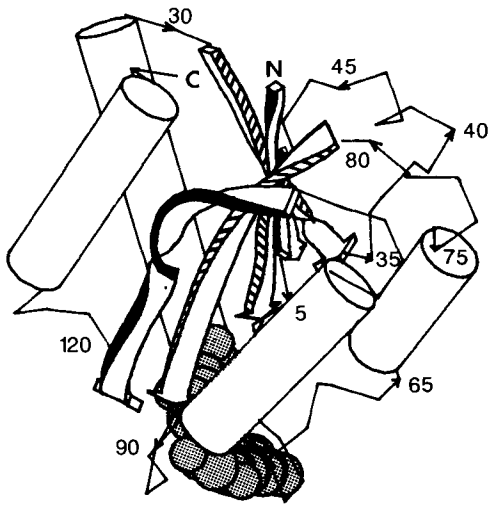

Fig. 1. *Clostridium* MP Flavodoxin structure. Picture produced by a computer program written by A.M. Lesk and K.D. Hardman.[37,38] Cylinders represent α-helices, ribbons represent β-sheets, and the FMN moiety is shown as space filling balls.

eter force field was used along with calculated hydrogen positions, no water molecules or FMN cofactor.

RESULTS

The structure of flavodoxin was predicted from only the α-carbon position and the directionality of the chain (Fig. 1). Since the X-ray structure of flavodoxin was known to 1.9 Å this provided an absolute check on the accuracy of the modeling procedure and of the usefulness of the benchmark parameters in assessing the validity of the model. The accuracy of the model was determined by calculating the RMSD from the observed X-ray structure for main-chain and side-chain atoms (Fig. 2) and comparing the dihedral angles ϕ, ψ, and the torsional χ_1, χ_2 angles (Fig. 3).

The secondary structure assignments used were those deposited with the crystal structure: Kabsch and Sander[17] assignments were also determined.

Main-Chain Atoms

The main chain was built to a RMSD of 0.51 Å across all backbone atoms, although when the outlier values considered below are omitted the RMSD was 0.43 Å. The experimental error of the flavodoxin X-ray atomic coordinate set can be estimated as 0.2 Å.[18] However, examination of cases where the same structure has been solved independently in different laboratories shows that errors are more likely to be approximately twice this value. The main-chain atoms of regular secondary elements, as defined by Kabsch and Sander, were modeled more accurately than the irregular regions: RMSD of α-helix = 0.36 Å, β-sheet = 0.55 Å, turns = 0.69 Å, and extended coil = 0.77 Å. The internal hydrogen bonding patterns of α-helix and β-sheet restrict the conformational possibilities of these backbone atoms and consequently such segments were predicted more accurately.

Five regions of the model had large RMSD (> 1.0 Å) and ϕ, ψ angles that deviated significantly from the X-ray structure; these were the residues 41–42, 57–58, 74–76, 89–92, and 117–121. In each region when the ϕ angle of residue i deviated from that of the native structure the ψ angle of residue $i+1$ was also displaced by roughly an equivalent amount but in the opposite direction. The effect of such a compensation on the geometry of the backbone is illustrated in Figure 4. The position of the side-chain Cβ atom was not significantly altered whereas the main-chain atoms N, C, and O were "flipped."

In four of the five poorly modeled regions error arises because the native X-ray structure was irregular. During the model-building procedure any residues that plotted into disallowed regions of the Ramachandran plot were reworked to bring the final ϕ, ψ values closer to idealized backbone parameters. In the native structure the residues Leu-43, Glu-59, Lys-76, Asn-119, and Asp-122 all plotted into disallowed regions. When the initial backbone was determined (Steps 1–3) the residues Glu-59, Ile-77, Asp-92, Val-111, Asn-119, and Glu-120 did plot into disallowed regions but they were reworked as outlined in Step 4. Four of these six "errors" were associated with, or close to, unusual regions in the native protein structure. The decision to rework the backbone to fit within well-defined ϕ, ψ values may have been in error.

The FMN moiety was not modeled into the structure as a knowledge of its binding site could not be predicted ab initio. Those residues that are within 4.0 Å of the FMN in the native structure are 7 to 12, 54 to 59, and 87 to 91. These residues were built with the same degree of precision as the remainder of the molecule: main-chain atoms = 0.69 Å, side-chain atoms = 2.30 Å, and all atoms = 1.57 Å. The Trp-90 side-chain atoms do penetrate volume normally occupied by the FMN moiety. However, this can be accounted for by a simple rotation around its χ_2 angle.

The segment (88-Tyr-Gly-Trp-Gly-Asp-Gly-Lys-Trp-95) was identified as an S turn by the Kabsch and Sander assignment (Fig. 4). One bifurcated hydrogen bond between Gly-89 and Gly-91 holds this segment to the FMN cofactor. The short peptide backbone dictionary approach to model building was unable to reproduce the geometry required for this particular element. This failure may be due to the unusually high glycine content of this short segment with its inherent high degree of flexibility. This segment may adapt its structure to the binding of the cofactor.

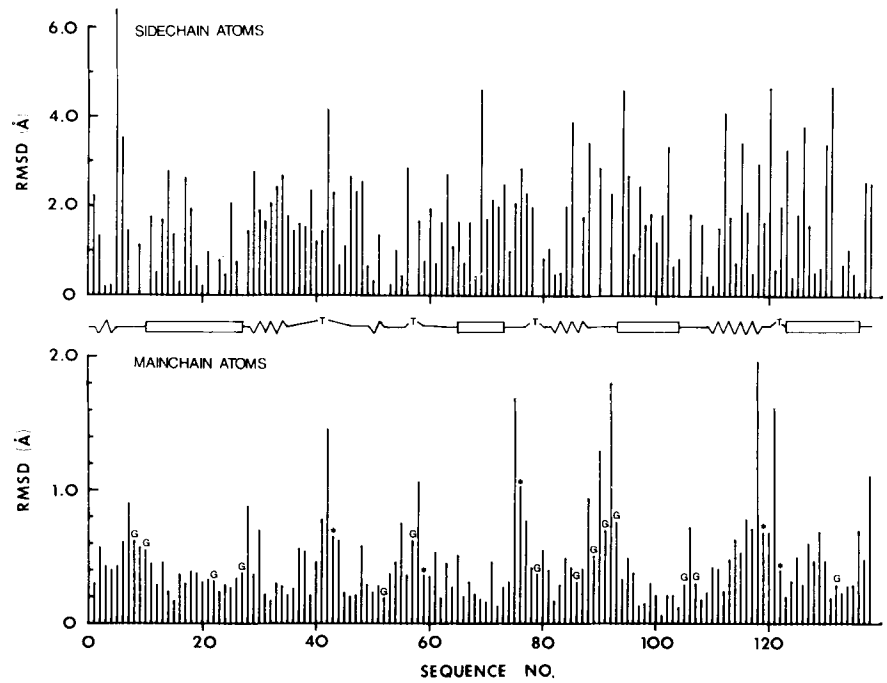

Fig. 2. RMSD of predicted model to X-ray crystal structure of flavodoxin. Main-chain residues of native FXN that map into disallowed regions of the Ramachandran plot are indicated by an asterisk, glycine residues by a G. The actual secondary structure is indicated as α-helix(—▭—); β-sheet (⩚); turn(–T–).

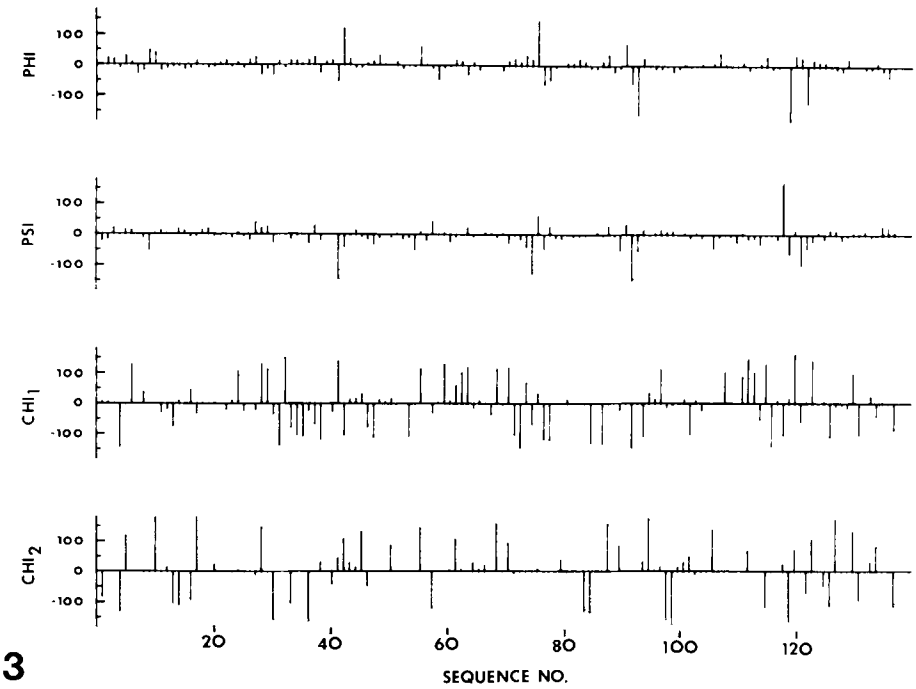

Fig. 3. Difference between ϕ, ψ, χ_1, χ_2 angles of the model FXN structure to the native X-ray structure.

Side-Chains

The side-chain atoms were built to a RMSD of 2.41 Å. The secondary elements were built with roughly the same degree of accuracy: RMSD of α-helix = 2.29 Å, β-sheet = 2.62 Å, turns = 2.19 Å, and extended coil = 2.51 Å.

A total of 25 buried residues, relative surface accessibility (<50%), had individual RMSDs of >1.5

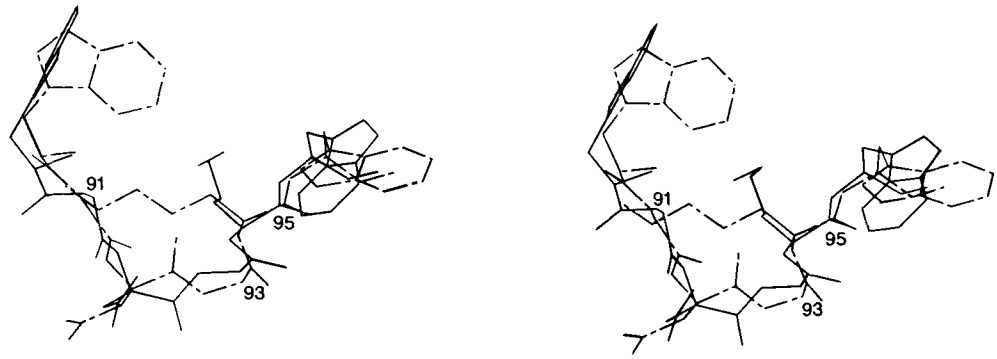

Fig. 4. Stereo diagram of the S turn region residues 90–95 of flavodoxin: model (– – –), X-ray (——). The peptide unit has been flipped at residue 92.

Å. Half of these poorly placed residues (12/25) were Val, Ile, and Leu and did not cluster in the core of the protein structure.

χ angle

The average population of the potential wells for side-chain χ_1 angles was determined for 75 proteins and for the native and model flavodoxin structures (Table I). In the model, the χ_1 angles were poorly predicted. On average, the χ_1 angle was wrong by 58° ± 53°. The χ_1 angles of the model were correct to within 20° of the X-ray values when set as g^+ = 67% (31/46), trans = 42% (13/31), g^- = 9% (1/11), or other = 7% (2/30) times. Thus, only 40% of the χ_1 angles, in total, was correctly located. The primary cause of error was due to placing the χ_1 angle in the g^+ position when it should have been trans (18 times) or the reserve (12 times). The g^- position was chosen six times for residues that occupy positions outside of the normal potential wells in the native structure. Positions outside of the normal potential wells were severely underpopulated because the initial placement of side chains was restricted to the three canonical positions.

The χ_2 angles were 30% correct (21/72 times) but only 17% (12/72) of all relevant residues had both the χ_1 and χ_2 angles correct. No significant distinction in the ability to predict torsional angles could be determined on the basis of relative exposure of the side chains in the model (data not shown).

Core residues

It was the initial placement of bulky side-chain residues, in particular, Ile, Trp, Phe, and Tyr within the "core" of the protein that determined the overall packing of the core. Figure 5A illustrates how the preliminary positioning of Phe-85 in the model occupied the volume normally taken up by Phe-131, which itself had subsequently moved out into solution. The hole left by Phe-85 was filled by Tyr-5. Thus, the initial placement error of Phe-85 was propogated throughout the core. On the other hand, the position of Phe-99 was correct and the volumes occupied by Phe-66 and Trp-95 were similar to the native structure (Fig. 5B). Once the bulky groups were positioned in the model small residues were easily built to fill the remaining space. Subsequent rotation of the bulky core residues encountered a high-energy barrier and initial placement was thus crucial to the overall packing strategy of the protein. A similar result has been recently reported by Bruccoleri and Karplus in their calculational approach to the modeling of side-chain residues.[19]

TABLE I. Distribution of the Side-Chain χ_1 Torsion Angle

Potential well		% Occupancy		
		Flavodoxin		
		Native	Model	Average*
g^+	−60° ± 20°	39.3	56.4	34.9
Trans	180° ± 20°	26.5	29.9	23.2
g−	60° ± 20°	9.4	6.0	12.9
Odd	All others	25.6	8.5	29.0

*From 75 proteins of resolution < 2.01 Å, giving 10,969 residues. These figures were derived using BIPED, the Birkbeck Protein Structure Oracle Database, which is funded by the U.K. Protein Engineering Club.

Pairs database

A database of 190,000 pairwise side-chain interactions was compiled and is summarized in Table II. As an example, 25 of the 26 Cys–Cys pairs found in the database that have superimposable main-chain atoms to the crambin (1CRN) Cys-16–Cys-26 disulfide bridge have the same side-chain χ_1 values (Fig. 6). It is clear that there is a dominant left-handed spiral configuration and, in the absence of other constraints, the bridge would be modeled in this manner. A similar approach has been advocated to the modeling of disulfide bonds by the program PROTEUS.[20]

A total of four side-chain interactions involving aromatic amino acids were ambiguous in the model

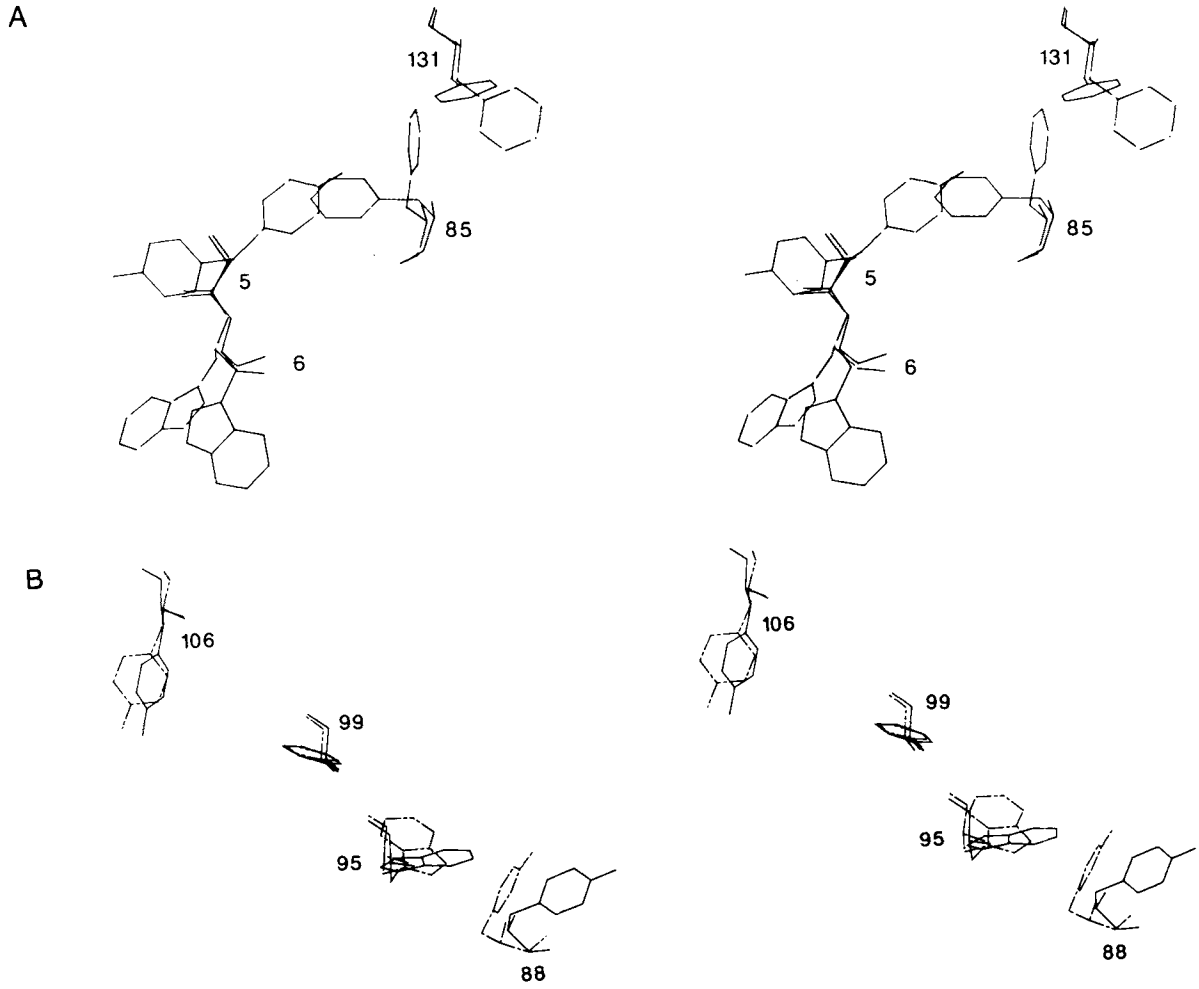

Fig. 5. Stereo diagram of aromatic core residues that adopt (**A**) an alternative conformation in the model or (**B**) maintain the internal X-ray structure packing motif: model (– – –), X-ray (——).

when only torsion angles were considered. A limited number of superimposable backbone geometries (within a 1.0 Å RMSD limit) were extracted from the PAIRS database. Results are summarized in Table III. An insufficient number of observations was found and it was not possible to orient these side chains with this technique. At this point, for other than disulfide bonds, the database still appears too small to be effective in real modelling situations.

Progress of the Refinement Procedure

During the rebuilding exercise the conformation of the model approached that of the native flavodoxin structure. The RMSD of the final energy minimized model to the X-ray crystal structure was 1.73 Å across all atoms of which 2.41 Å was from the side-chain atoms and 0.57 Å from the main-chain atoms. Figure 7 indicates the RMSD at each intermediate state for main-chain atoms and also the contribution due to the side-chain atoms.

The greatest improvement occurred when van der Waals clashes involving side-chain atoms were resolved. A total of 16 clashes were found at the main-chain plus Cβ level (Step 5). After all of the side-chain torsional angles were set to their statistically optimal position a total of 37 new clashes were found. All of the side-chain clashes could be resolved by choosing new torsional angles in accordance with statistical profiles: that is, no atoms were closer than 2.4 Å. Some contacts were found in the region of 2.4–2.5 Å but it was decided to defer the resolution of such conflicts until the energy minimization step.

In the final step of the procedure a conjugate gradient, Cartesian-space energy minimizer was utilized. The minimization procedure did not move the model toward the native structure even though it did locate a local potential energy well. The overall structure did not change significantly (all atom RMSD = 0.52 Å): the surface area of the model decreased by 6.5% during the minimization procedure. However, this change is due primarily to the move-

TABLE II. Number of Pair-Wise Side-Chain Interactions in the PAIRS Database*

	A	N	D	R	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	671																			
N	647	397																		
D	704	451	166																	
R	366	334	408	112																
C	432	296	211	148	577															
Q	653	430	246	243	228	137														
E	743	431	363	347	139	318	228													
G	950	540	688	360	442	553	383	573												
H	369	275	392	196	166	187	295	377	179											
I	904	404	470	275	302	373	514	618	263	441										
L	1431	756	544	465	483	606	680	1061	550	1168	978									
K	1083	676	748	269	324	373	784	727	406	667	859	364								
M	293	167	161	109	113	132	127	169	105	245	300	183	41							
F	688	354	368	372	274	331	375	416	448	660	1185	569	201	412						
P	563	323	295	238	157	385	313	396	247	277	650	386	95	333	133					
S	922	608	612	348	556	515	530	816	397	717	1038	847	205	470	533	495				
T	895	571	583	371	275	429	395	781	259	612	876	672	171	593	474	848	282			
W	292	289	225	174	130	228	181	270	121	397	591	315	128	279	191	335	219	88		
Y	672	583	433	312	344	335	309	616	315	634	675	653	203	439	425	629	467	251	243	
V	1354	567	511	326	370	573	593	1145	482	1043	1820	847	303	874	543	970	790	504	760	695
	Total number of interactions per residue type																			
	14678	9119	8595	5782	5971	7283	8056	11909	6039	11018	16749	11776	3459	9666	6964	12420	10578	5213	9307	15098
	Total number of residues																			
	2860	1645	1641	872	992	1156	1450	2869	832	1581	2426	2176	434	1203	1333	2767	2003	555	1124	2423
	Average number of interactions per residue																			
	5.13	5.54	5.24	6.63	6.02	6.30	5.56	4.15	7.26	6.97	6.90	5.41	7.97	8.03	5.22	4.49	5.28	9.39	8.28	6.23

Overall average number of interactions: (189680 / 32342) = 5.86

*Calculated on a database of 159 proteins of resolution < 2.5 Å.

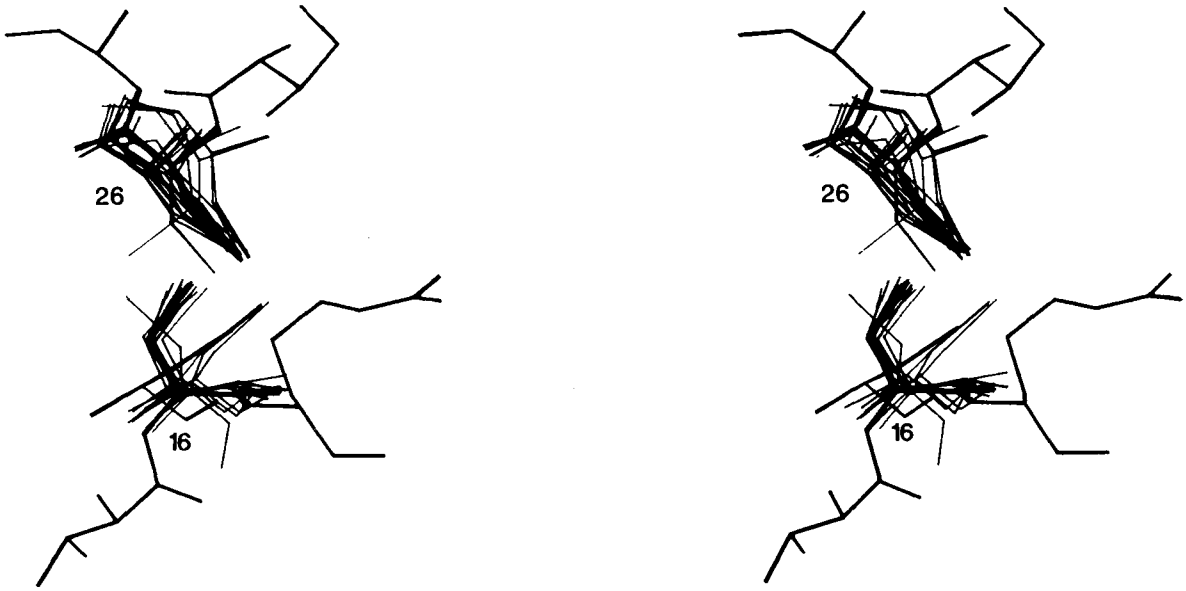

Fig. 6. Superposition of all the 26 disulfide backbone pairs found in the side-chain database (thin lines) that fit to within 1.0 Å RMSD of the crambin Cys-16–Cys-26 pair (thick line).

TABLE III. Validity of Side-Chain Guide Conformations Extracted From the PAIRS Database for Four Pairs of Ambiguous Side-Chain Interactions in the FXN Model*

	Residue I χ_1			Residue J χ_1	
	Correct†	Total‡		Correct	Total
Ile-3	19	23	Tyr-5	5	23
Tyr-5	5	6	Gly-52	NA§	NA
Phe-66	2	2	Phe-99	2	2
Phe-85	0	6	Phe	3	6

*Search procedure: (1) find all pairs where the Cα–Cα distance is ±1.0 Å; (2) superimpose the first residue of the test pair onto the main-chain atoms of residue I; (3) reject the test pair if RMSD > 1.5 Å across all atoms.
†The guide χ_1 angle agrees to the native FXN χ_1 by ±20°.
‡Total number of pairs found in the database.
§NA, not applicable.

ment of side-chain atoms as the surface area associated with the backbone atoms decreased by only 1.4% (data not shown). Native flavodoxin (no FMN, no water) was also minimized under the same conditions as the model. This caused an increased divergence in the two structures: RMSD of minimized model to the minimized flavodoxin across all atoms = 1.86 Å, side-chain atoms = 2.45 Å, and main-chain atoms = 0.51 Å.

Test Criteria and Model Analysis Procedure

For assessment of the model several benchmark parameters were calculated from known protein crystal structures of resolution better than 1.6 Å. The proteins used were trypsin inhibitor (5PTI), rubredoxin (5RXN), neurotoxin (1NXB), avian pancreatic polypeptide (1PPT), erythrocruorin (1ECA), myoglobin (1MBN), ribonuclease A (1RN3), β-trypsin (1TPP), cytochrome *c* [rice] (1CCR), crambin (1CRN), insulin (1INS), lysozyme (1LZ1), proteinase A (2SGA), staphylococcal nuclease (2SNS), cytochrome *c* [tuna] (4CYT), and carboxypeptidase A (5CPA).

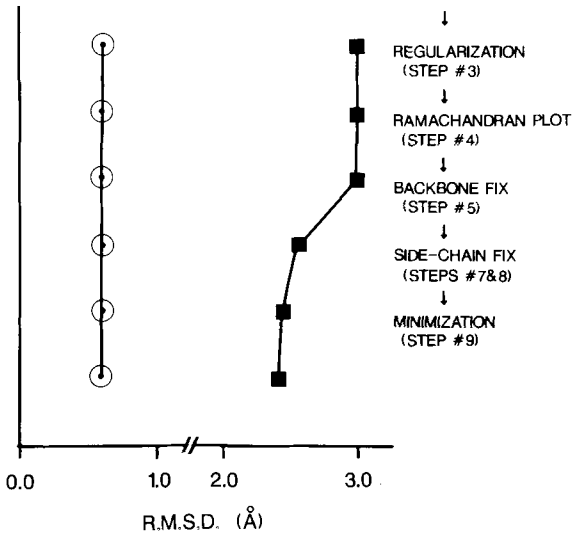

Fig. 7. Improvement in the RMSD error of the model FXN structure at various stages during the modeling procedure: side-chain atoms (■——■); main-chain atoms (○——○).

Surface area

The accessible surface area (not the contact area) of the model to a 1.4 Å spherical probe (equivalent to

TABLE IV. The Distribution of Relative Side-Chain Accesibilities in High-Resolution Proteins to a Water Molecule*

Residue	N	% Accessibility well†										
		10	20	30	40	50	60	70	80	90	100	>100
Ala	957	43.4	8.7	6.3	5.1	4.2	4.6	4.8	4.9	5.2	3.7	9.2
Arg	308	17.9	8.4	12.0	11.0	10.4	9.7	11.4	9.7	5.2	1.6	2.6
Asn	533	19.9	6.6	11.1	7.1	11.3	7.1	9.2	11.8	7.1	4.3	4.5
Asp	665	20.8	6.3	6.0	6.8	9.9	9.9	8.4	9.5	5.7	6.5	10.2
Cys	335	73.1	8.4	4.5	4.2	2.1	2.1	2.1	0.9	0.9	0.9	0.9
Gln	415	21.4	4.6	8.9	9.2	10.8	9.9	14.9	7.2	7.2	3.9	1.9
Glu	525	17.3	4.8	7.6	7.6	8.0	11.0	7.8	12.2	9.7	7.2	6.7
His	313	46.0	8.6	6.1	7.7	4.5	6.1	7.6	6.7	3.2	1.9	1.6
Ile	526	61.6	7.4	7.2	6.7	4.9	3.8	2.7	2.9	1.0	1.3	0.6
Leu	819	61.2	8.9	7.1	7.6	4.4	3.5	1.6	1.9	1.6	0.9	1.3
Lys	761	12.9	3.2	3.8	7.5	14.2	14.5	13.5	12.4	9.1	5.9	3.2
Met	152	62.5	8.6	3.9	5.3	3.9	3.3	6.6	2.6	2.0	0.0	1.3
Phe	426	60.1	12.4	7.1	4.7	4.7	3.1	3.3	2.1	0.5	1.4	0.7
Pro	514	27.4	8.4	6.0	11.5	11.3	10.1	11.3	8.4	4.9	0.8	0.0
Ser	940	26.7	5.7	6.4	4.7	8.7	7.6	7.9	8.5	8.8	6.4	8.6
Thr	776	28.1	5.8	8.2	10.6	11.0	9.0	9.3	8.1	4.5	3.0	2.4
Trp	181	48.6	21.0	9.4	5.5	3.9	2.2	2.8	5.1	1.1	0.0	0.0
Tyr	425	40.7	15.8	13.2	8.2	7.3	4.2	4.7	2.8	0.9	1.6	0.5
Val	854	59.9	9.1	7.3	5.9	3.6	4.9	2.7	3.3	2.1	0.8	0.4
a	5651	24.1	6.4	7.8	7.7	9.8	9.2	9.5	9.2	6.6	4.7	4.8
b	4765	54.1	9.4	6.7	6.4	4.8	4.5	4.0	3.7	2.5	1.5	2.4
c	10416	37.8	7.8	7.3	7.1	7.5	7.1	7.0	6.7	4.8	3.2	3.7

*N, number of residues in the database. a, hydrophilic residues: Arg, Asn, Asp, Gln, Glu, His, Lys, Ser, Thr, Tyr. b, hydrophobic residues: Ala, Cys, Ile, Leu, Met, Phe, Pro, Trp. c, all residues.
†Relative to the area associated with the tripeptide Gly-X-Gly where $\phi = -139°$, $\psi = 135°$, $\chi_1 = 120°$, and all other torsional angles = 180°.

a water molecule) was calculated[21] and compared to average values observed for the same residues in equivalent secondary structural elements. The relative side-chain accessibilities of the test set of high resolution proteins was determined; results are given in Table IV.

Hydrophobic residues, as expected, were primarily buried but a significant proportion of the atoms of polar and charged side-chains had a relative accessibility of less than 10%; hydrogen bonding often accompanies the internalization of polar atoms.[22,23] The only real exception was that residues that were located in turns were more highly exposed; this was reflective of turns being found primarily on the surface of proteins.[24]

On average, the distribution of accessibilities was 6.2% of the total in each well, excluding the 0–10% well. Residues in the model were flagged when they fell into a well with less than 7% occupancy (Fig. 8). Such residues were about equally divided between exposed hydrophobics and buried hydrophilics. Only three of these buried hydrophilic residues (Lys-2, Ser-54, Ser-74) did not have a hydrogen bond to the side-chain moiety.

Internal cavity volume

Cavities (unfilled volume that does not communicate with the exterior of the protein and may indicate incorrect side-chain packing) were identified by the method of Moult.[25] Protein coordinates (no hydrogen atoms) were projected onto a 0.6 Å three-dimensional grid, and the volume inaccessible to a 1.4 Å spherical probe surrounding each atom was determined. Cavity volumes calculated from the test set of high-resolution proteins are shown in Table V. No significant difference was noted when cavity volumes were analyzed on the basis of secondary structure (data not shown).

The difference in cavity volume around each residue to the test protein values is shown in Figure 8. The volume of all buried atoms in the model is 4865 $Å^3$ as opposed to 5470 $Å^3$ for flavodoxin without the FMN moiety (or 5550 $Å^3$ for the whole enzyme). The average standard deviation from the test parameters for the model FXN is $-0.40 \pm 1.08\ \sigma$.

Close contact distances

The average number of residues that surrounded a particular residue was calculated in the following manner. If any atom of residue i was within 4.0 Å of any atom of a nonadjacent residue j then this was counted as a hit. In the case of adjacent residues only the side-chain atoms were considered in the calculation. Residues at all levels of solvent accessibility were used and this factor is therefore reflected in the magnitude of the standard deviation.

When surface exposed and buried residues are considered together the number of residues in contact with any other residue is primarily related to the volume the amino acid occupies rather than its

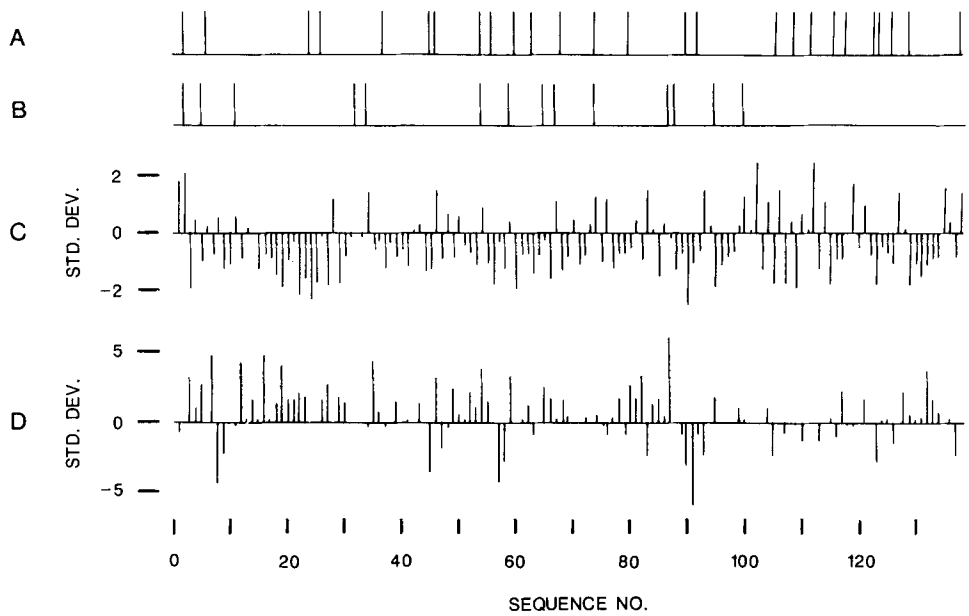

Fig. 8. Identification of residues in the model flavodoxin structure that fail various packing tests. **A:** side-chain atoms are flagged that had an accessibility occupied by < 7% of the same residue type in high-resolution proteins. **B:** polar side-chain atoms are flagged that have less than 50% relative accessibility to water and no hydrogen bonds. **C:** deviation of residue cavity volume from the average determined for the test set of proteins. Units are in standard deviations. **D:** deviation of residue contact number from the average determined for the test set of proteins. Units are in standard deviations.

TABLE V. The Distribution of Cavity Volumes in High-Resolution Proteins*

		All atoms		Main-chain atoms		Side-chain atoms		Polar atoms		Nonpolar atoms	
Residue	*N*	(Å^3)	σ	(Å^3)	σ	(Å^3)	σ	(Å^3)	σ	(Å^3)	σ
Ala	159	24.7	(6.5)	16.0	(3.9)	8.8	(3.8)	13.2	(3.4)	11.5	(4.3)
Arg	70	50.4	(10.8)	16.4	(4.0)	34.0	(8.6)	33.6	(8.6)	16.8	(3.8)
Asn	107	36.7	(9.2)	15.1	(4.3)	21.6	(6.6)	28.8	(8.7)	7.9	(3.8)
Asp	82	37.1	(9.2)	15.4	(3.6)	21.7	(7.4)	29.6	(7.7)	7.5	(2.4)
Cys	74	28.6	(5.3)	16.0	(3.6)	12.6	(3.3)	13.5	(3.3)	15.1	(3.6)
Gln	76	41.9	(9.0)	15.5	(3.7)	26.4	(6.8)	29.0	(8.8)	13.0	(5.3)
Glu	91	40.2	(9.4)	15.2	(3.8)	25.0	(7.1)	28.2	(7.7)	12.0	(3.3)
Gly	180	18.1	(5.2)	18.1	(5.2)	NA		13.2	(4.3)	4.9	(1.8)
His	50	43.0	(8.8)	15.1	(3.3)	28.0	(7.5)	22.8	(4.4)	20.2	(5.8)
Ile	102	48.9	(10.1)	16.5	(3.5)	32.4	(8.9)	14.2	(3.3)	34.7	(9.1)
Leu	126	51.0	(9.7)	16.9	(3.5)	34.1	(8.3)	14.3	(3.5)	36.8	(8.4)
Lys	139	38.4	(7.2)	15.6	(3.9)	22.7	(5.3)	18.4	(4.3)	20.0	(4.8)
Met	28	50.5	(9.1)	18.0	(3.8)	32.5	(8.1)	15.1	(3.6)	35.3	(8.1)
Phe	75	63.5	(12.8)	17.0	(4.2)	46.5	(11.3)	14.2	(3.9)	49.2	(11.8)
Pro	79	27.4	(7.7)	13.6	(4.2)	13.8	(4.8)	11.3	(3.7)	16.1	(5.3)
Ser	156	26.8	(7.6)	14.8	(4.2)	12.0	(5.1)	20.7	(6.4)	6.1	(2.3)
Thr	134	34.0	(8.6)	14.9	(3.8)	19.1	(6.5)	21.9	(5.7)	12.1	(4.2)
Trp	26	73.0	(12.3)	15.8	(3.5)	57.1	(11.6)	20.2	(3.9)	52.8	(11.0)
Tyr	94	58.3	(11.4)	15.9	(3.6)	42.4	(10.0)	23.8	(5.8)	34.5	(7.3)
Val	119	40.5	(8.7)	17.1	(3.5)	23.4	(7.3)	14.4	(3.2)	26.1	(7.4)

*N, number of residues in database; σ, standard deviation; NA, not applicable.

hydrophobicity.[26] The number of contacts increases linearily with an interatomic cutoff distance over the range of 2.5–8 Å.

During the "hand building" stages of the modeling procedure (Steps 1–8) atoms within the van der Waals contact limit (2.4 Å) were moved to just beyond it. Any interatomic bad contacts were relieved by the subsequent energy minimization step. At the end of this procedure residues were still too tightly packed, as shown by the high number of contacts found over the cutoff distance range of 2.5–3.5 Å. Application of a conformational search, side-chain torsion angle driver energy minimization procedure could be used to "even out" the packing.

TABLE VI. Energy Minimization of Model Built and Native Flavodoxin Structures*

Flavodoxin†	Start‡	Finish§
Native		
Bonded	472	397
Coulombic	−2734	−3160
van der Waals	−1276	−1305
Total	−3538 kcal	−4068 kcal
RMSD atomic shift	0.17 Å	0.43 Å
Model		
Bonded	594	393
Coulombic	−2434	−3063
van der Waals	−1138	−1225
Total	−2977 kcal	−3895 kcal
RMSD atomic shift	0.19 Å	0.52 Å

*Calculated on 1324 bonds, 1913 angles, 2549 torsion angles, 516 wags, 138 constraining Cα atoms, and 1308 nonbonded atoms.
†FMN and water not included.
‡After 100 cycles.
§After 600 cycles.

Hydrogen bonding patterns

The HBOND option of the graphics program HYDRA[27] was used to calculate all hydrogen bonds in the model. Correct bonds were defined as an O–N distance of less than 3.3 Å, and the three angles involved ≥ 90° (X-Donor-H-Acceptor-X). This test identified poor backbone geometries in α-helical and β-sheet secondary structures. Also, this procedure identified, by omission, those polar atoms that were buried with no possibility of making a hydrogen bond.

The lack of hydrogen bonds to a buried polar or charged group is a strong indicator of a potentially misfolded side chain.[28,29] Those polar residues which were buried (<50% polar side-chain exposure) and lacked a hydrogen bond are shown in Figure 8.

Energy criteria

An energy minimization procedure was applied to both the native flavodoxin and the model; results are summarized in Table VI. After minimization the model had a better energy than that of the starting native structure although the minimized native structure was significantly more stable than the final model. Clearly, it was possible to create a protein model that, while energetically satisfactory, was packed with a substantially different core conformation.

A direct comparison between the model and X-ray structure minimization profiles may be suspect. The structures dropped significantly in total energy but with a relatively low atomic coordinate RMSD. The structures did not contain water or the flavodoxin prosthetic group, which would affect the relative energy values. The important point, however, was that with only a minimal change to the overall structure a low energy form was obtained for the model compound. This was in keeping with the work of Novotny in which it was shown that deliberately misfolded structures could minimize to a stable but nonnative structure.[30]

Overall assessment

The primary finding, as outlined in Table VII, was that outliers, to any particular test, exist in any real protein structure. Consequently, one assessment parameter was not enough to identify incorrectly packed residues.

Several bad assessments may be indicative of a problem. However, in this case, when the four tests discussed above were taken together or in various combinations this was still insufficient to identify a series of residues that were packed incorrectly relatively to the crystal structure.

DISCUSSION

This work was undertaken to determine the accuracy with which the three-dimensional structure of a protein could be predicted with knowledge based techniques given only a "perfect" Cα trace. Temperature factors have not been deposited for the flavodoxin structure and as a consequence, all residues in the "perfect" native conformation were assumed to

TABLE VII. Analysis of Error Associated With Subsets of Residues Identified as Incorrectly Placed

	RMSD*					
	Side-chain atoms		Main-chain atoms		All atoms	
Testing procedure	(Å)	N†	(A)	N	(A)	N
Accessibility	2.54	115	0.74	130	1.92	245
Cavities	2.37	235	0.56	251	1.73	506
Close contacts	2.38	365	0.58	412	1.69	777
Hydrogen bonds	2.98	68	0.46	56	2.22	124
All tests‡	1.00	2	0.31	4	0.53	6
All atoms§	2.41	520	0.57	552	1.73	1072

*RMSD is of the model to native FXN.
†N, number of atoms.
‡Only Ser-54 failed each of the four testing procedures.
§Accuracy of all atoms in the entire model to the native structure.

have equal validity. To assume that any one X-ray structure will lie completely within the statistical average is clearly wrong. The specific conformations of the native structure that deviate from the average values are reflective of events associated with the folding pathway of the protein. This folding history may trap specific residues in higher potential wells in an effort to speed the folding process or to allow certain secondary structures to form.[31] In addition, the subsequent interaction with cofactors may slightly alter the conformation of the backbone and certainly will affect the position of the side-chain atoms.

There is a significant degree of subjectivity in traditional protein modeling techniques and this work attempts to address the degree to which models can be trusted. Only recently have programs such as COMPOSER[32] become available for the automated task of model building highly homologous proteins. The modeling procedure, which has no information as to folding pathway or cofactor placement, attempts to place residues in the best possible conformation from an energetic point of view. Consequently, a model could be created with a energy lower than that of the native protein. Indeed, additional efforts to improve the flavodoxin model may well lower its energy below that of the native structure.

In the first step, the backbone was modeled from fragments of known structures. This was broadly successful with only five short segments in error. Most of these "errors" correspond to a flip of the peptide unit that had disallowed ϕ, ψ values in the native protein.

An alternative backbone building procedure, which was not applied in this work, may be to follow the procedure of Purisima and Scheraga.[33] In this approach an idealized backbone is generated mathematically given only the Cα coordinates and one pair of ϕ, ψ angles. These angles are usually generated from an internal proline residue. This method does suffer from the restriction that the Cα coordinates be at close to an ideal separation distance (3.8 Å) but it can produce a backbone with a RMSD of = 0.3 Å to the native structure, a result similar to this work. This method could be used to cross-check and thus identify particular residues that were "flipped."

The dictionary approach, however, is still to be favored as the initial building procedure for three reasons. First, a range of backbone conformations that can fit a particular Cα trace are clearly indicated on the graphic display. The presence of "families" in the backbone structures was apparent. These traces also include examples where residues have ψ, ϕ values that plot outside of well-defined regions of the Ramachandran plot. Second, the procedure can easily handle Cα coordinates that are at less than ideal separation distances or angles. And finally, an option exists to allow certain Cα atoms to be unconstrained during the search. Judicious use of this option established the consistency of the proposed backbone structure in the region under analysis.

The modeling of side-chain conformations was much less successful. It was hoped that knowledge of side-chain interactions and conformations, as compiled from statistical analysis of well-refined structures, would have predictive power for the de novo placement of side chains. This procedure failed for two reasons. First, the χ angle preference ratios were too low; i.e., local packing constraints do affect the positioning of side chains. Second, the database of side-chain interactions was too small to be of general value.

The placement of residues, according to statistical averages, is best assessed by the results obtained with the core residues. Only some of the critical orientations of large hydrophobic residues could be determined from statistical methods. Work by Zehfus[34] and many others on the presence of contiguous compact regions may help to identify those residues that begin the folding process and potentially pack together. However, the usefulness of such information in predictive schemes is unknown.

Alternatively, to attempt to position all side-chain conformations by dynamic calculational procedures would be impossible with current techniques. For example, the distance across which a residue can influence its neighbors is an important but poorly understood variable. In the case of single-site mutations the new residue is often accommodated by short-range fluctuations in the spatially adjacent residues. However, significant long-range conformational rearrangements have also been found.[35] Finally, the computer time required to search out even a restricted subset of the possible configurations of all the interacting core residues would certainly be too prohibitive to make this a general procedure.

Even though models can be generated that are realistic, when considered in their totality, it is vital to be able to identify regions or particular residues that have been incorrectly packed. Quantitative parameters were measured for high-resolution structures and statistically generated values utilized in the validation of the model. The tests, however, did not identify those residues that had a conformation different from the X-ray structure; they identified only those side chains with a conformation that differed from the statistical average. Also, the test results did not specifically indicate how to alter the side-chain conformations to bring them into line with the statistical averages.

It may be possible to improve the placement of residues that have been identified as suspect by adapting the rotamer library concept of Ponder and Richards[36] and Sternberg[12] into a conformational search procedure whereby large aromatic residues

are initially constrained to a favorable torsional angle subset based on their secondary structure. Nearby groups would then be given greater flexibility to search through available space although additional weight would be given to preferred torsional angles.

The database of pairwise side-chain interactions, described above, is a second approach to this rebuilding problem. This methodology does not explicitly account for the influence of specific nearby side chains that probably affect the packing in any real example. However, the examples found do represent real ways of packing two residues together in the presence of other side chains and thus may be of use.

If the crystal structure is available PAIRS may also be of use to predict the suitability of a proposed single-site amino acid replacement experiment. However, given the constraint that the backbone atoms must be superimposable in real space the actual number of entries that is found to align is often not statistically significant for other than disulfide bridges. Unlike energy minimization or folding rule methodologies this technique does have the inherent advantage that each new crystal structure will add $6N$ (N = number of residues) pairwise interactions to the database.

ACKNOWLEDGMENTS

LSR held a MRC (Canada) postdoctoral fellowship. We thank the SERC for financial support. The authors thank Tim Hubbard for assistance with the internal cavities calculation, Dr. I. Haneef for assistance with the energy minimization calculation, and Prof. Tom Blundell for helpful discussions.

REFERENCES

1. Cohen, F.E., Kuntz, I.D. Prediction of the three-dimensional structure of human growth hormone. Proteins 2: 162–166, 1987.
2. Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E., Thornton, J.M. Knowledge-based prediction of protein structures and the design of novel molecules. Nature (London) 326: 347–352, 1987.
3. Hagler, A.T., Honig, B. On the formation of protein tertiary structure on a computer. Proc. Natl. Acad. Sci. U.S.A. 75:554–558, 1978.
4. Chothia, C., Lesk, A.M. The relation between the divergence of sequence and structure in proteins. EMBO J. 5: 823–826, 1986.
5. Hubbard, T.J.P., Blundell, T.L. Comparison of solvent inaccessible cores of homologus proteins: Definitions useful for protein modelling. Protein Engineer. 1:159–171, 1987.
6. Chothia, C., Lesk, A.M., Levitt, M., Amit, A.G., Maiuzza, R.A., Phillips, S.E.V., Poljak, R.J. The predicted structure of immunoglobulin D 1.3 and its comparison with the crystal structure. Science 233:755–758, 1986.
7. Smith, W.W., Burnett, R.M., Darling, G.D., Ludwig, M.L. Structure of the semiquinone form of flavodoxin from *clostridium* MP. Extension of 1.8 Å and some comparisons with the oxidized state. J. Mol. Biol. 117:195–226, 1977.
8. Summers, N.L., Carlson, W.D., Karplus, M. An analysis of sidechain orientations in homologous proteins. J. Mol. Biol. 196:175–198, 1987.
9. Novotny, J., Rashin, A.A., Bruccoleri, R.E. Criteria that discriminate between native proteins and incorrectly folded models. Proteins 4:19–30, 1988.
10. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B. Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
11. Jones, T.A., Thirup, S. Using known substructures in protein model building and crystallography. EMBO J. 5:819–822, 1986.
12. McGregor, M.J., Islam, S.A., Sternberg, M.J.E. Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. J. Mol. Biol. 198: 295–310, 1987.
13. Reid, K.S.C., Lindley, P.F., Thornton, J.M. Sulphur-aromatic interactions in proteins. FEBS Lett. 190:209–213, 1985.
14. Sawyer, L., James, M.N.G. Carboxyl-carboxylate interactions in proteins. Nature (London) 295:79–80, 1982.
15. Singh, J., Thornton, J.M. The interaction between phenylalanine rings in proteins. FEBS Lett. 191:1–5, 1985.
16. Haneef, I. University of London, Ph.D. thesis, 1986.
17. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
18. Luzzati, P.V. Traitement statistique des erreures dans la determination des structures cristallines. Acta Crystallogr. 5:802–810, 1952.
19. Bruccoleri, R.E., Karplus, M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 26:137–168, 1987.
20. Pabo, C.O., Suchanek, E.G. Computer-aided model-building strategies for protein design. Biochemistry 25: 5987–5991, 1986.
21. Connolly, M. Computation of molecular volume. J. Am. Chem. Soc. 107:1118–1124, 1985.
22. Quiocho, F.A., Sack, J.S., Vyas, N.K. Stabilization of charges on isolated groups sequestered in proteins by polarized peptide units. Nature (London) 329:561–564, 1987.
23. Lawrence, C., Auger, I., Mannella, C. Distribution of accessible surfaces of amino acids in globular proteins. Proteins 2:153–161, 1987.
24. Wolfenden, R., Andersson, L., Cullis, P.M., Southgate, C.C.B. Affinities of amino-acid side chains for solvent water. Biochemistry 20:849–852, 1981.
25. Moult, J. Personal communication, 1987.
26. Reid, L.S., Thornton, J.M. Prediction of protein structure from Cα atomic coordinates. UCLA Symposium NS 69, 93–103, 1987.
27. HYDRA is a real-time graphics display program distributed by POLYGEN corporation (Waltham, MA).
28. Baker, E.N., Hubbard, R.E. Hydrogen bonding in globular proteins. Prog. Biophys. Molec. Biol. 44:97–179, 1984.
29. Chothia, C. Principles that determine the structure of proteins. Annu. Rev. Biochem. 53:537–572, 1984.
30. Novotny, J., Bruccoleri, R., Karplus, M. An analysis of incorrectly folded protein models: Implications for structural predictions. J. Mol. Biol. 177:787–818, 1984.
31. Wetlaufer, D. Prolyl isomerization: How significant for *in vivo* protein folding? Biopolymers 24:251–255, 1985.
32. Blundell, T., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D.A., Sibanda, B.L., Sutcliffe, M. Knowledge-based protein modelling and design. Eur. J. Biochem. 172:1179–1188, 1988.
33. Purisima, E.O., Scheraga, H.A. Conversion from a virtual-bond chain to a complete polypeptide backbone chain. Biopolymers 23:1207–1224, 1984.
34. Zehfus, M.H. Continuous compact protein domains. Proteins 2:90–110, 1987.
35. Brayer, G. Personal communication, 1987.
36. Ponder, J.W., Richards, F.M. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:775–791, 1987.
37. Lesk, A.M., Hardman, K.D. Computer-generated schematic diagrams of protein structures. Science 216:539–540, 1982.
38. Lesk, A.M., Hardman, K.D. Computer-generated pictures of proteins. Methods Enzymol. 115:381–390, 1985.