Journal of Computer-Aided Molecular Design, 7 (1993) 199–224
ESCOM

J-CAMD 191

PROGEN: An automated modelling algorithm for the generation of complete protein structures from the α-carbon atomic coordinates

Chhabinath Mandal and D. Scott Linthicum*

Center for Macromolecular Design and the Department of Veterinary Pathobiology, College of Veterinary Medicine, Texas A&M University, College Station, TX 77843-4467, U.S.A.

Received 22 June 1992
Accepted 6 November 1992

Key words: Polypeptide; Computer-aided modelling; Protein database; Optimal geometry parameters

SUMMARY

A modelling algorithm (PROGEN) for the generation of complete protein atomic coordinates from only the α-carbon coordinates is described. PROGEN utilizes an optimal geometry parameter (OGP) database for the positioning of atoms for each amino acid of the polypeptide model. The OGP database was established by examining the statistical correlations between 23 different intra-peptide and inter-peptide geometric parameters relative to the α-carbon distances for each amino acid in a library of 19 known proteins from the Brookhaven Protein Database (BPDB). The OGP files for specific amino acids and peptides were used to generate the atomic positions, with respect to α-carbons, for main-chain and side-chain atoms in the modelled structure. Refinement of the initial model was accomplished using energy minimization (EM) and molecular dynamics techniques. PROGEN was tested using 60 known proteins in the BPDB, representing a wide spectrum of primary and secondary structures. Comparison between PROGEN models and BPDB crystal reference structures gave r.m.s.d. values for peptide main-chain atoms between 0.29 and 0.76 Å, with a grand average of 0.53 Å for all 60 models. The r.m.s.d. for all non-hydrogen atoms ranged between 1.44 and 1.93 Å for the 60 polypeptide models. PROGEN was also able to make the correct assignment of *cis*- or *trans*-proline configurations in the protein structures examined. PROGEN offers a fully automatic building and refinement procedure and requires no special or specific structural considerations for the protein to be modelled.

INTRODUCTION

In the past few years the collection of detailed structural information for proteins has been growing in an exponential-like fashion. Determinations of protein structures by X-ray crystallography and NMR spectroscopy have contributed significantly to this expanding database. Much of the crystallographic information for proteins is compiled in the Brookhaven Protein Databank

*To whom correspondence should be addressed.

(BPDB) and this databank has proved valuable to a wide variety of investigations on protein structure and function [1]. For some protein crystals, however, the data entries in the BPDB have been limited to only α-carbon coordinates. The need to refine the coordinates for all atoms can often delay the complete coordinate entries.

Several investigators have developed different computer-aided model building algorithms for the purposes of constructing all atom structures from incomplete (C^{α} only) coordinates or to be used as a tool for refinement of experimentally derived determinations [2–5]. Such modelling algorithms may also prove useful for the positioning of polypeptide side-chain atoms which lack high resolution coordinates. Reid and Thornton [2] developed a method which utilized a 'dictionary' of structural templates from known structures for the generation of main-chain atomic positions. Side-chain conformations were defined by statistical distributions, pair-wise side-chain interactions, and energy minimization (EM) techniques. As a test case, flavodoxin (3FXN) was rebuilt from C^{α} coordinates; a peptide backbone root-mean-square deviation (r.m.s.d.) of 0.6 Å and all atom r.m.s.d. of 1.7 Å were achieved [3]. Correa [4] developed an automatic building algorithm which generated a peptide backbone using only alanines, glycines, and prolines on a C^{α} coordinate scaffolding. The second stage of the procedure involved side-chain construction and refinement using molecular dynamics. For α-lytic protease (2ALP) a backbone r.m.s.d. of 0.19 Å and overall r.m.s.d. of 1.24 Å were attained. Model building of all atoms coordinates has also been achieved [5] using a series of peptide fragments which fit the α-carbon distances; side chains were added using a library frequently occurring rotamers, followed by a Monte Carlo procedure with simulated annealing. This procedure had an accuracy of 1.6 Å (r.m.s.d.) for positions of core side-chain atoms and 70% of the χ_1 angles were within 30° of the crystal structure values [5].

Some investigators might benefit from a complete, but not definitive structure, which could reveal areas of important secondary structure, solvent accessible residues and internal packing regions. Our work on antigenic determinants [6] and surface architecture of two plant proteins, thaumatin and monellin, which possess intense sweet taste properties [7], requires knowledge of the complete structure, but at the present time only the α-carbon coordinates are on deposit in the BPDB [8,9]. Thus, we were stimulated to create an automatic and accurate modelling algorithm which could GENerate the complete PROtein structure (PROGEN). PROGEN uses an optimal geometry parameter (OGP) database established by correlating the various intra-peptide and inter-peptide structural parameters (n = 23) and their relationship with C^{α} distances for each amino acid; this OGP database was established using 19 known protein structures available in the BPDB. PROGEN uses a statistical approach in generating the model atomic positions for individual atoms based on the OGP files. The initial generation of the model peptide backbone was followed by refinement procedures involving EM and dynamics simulations. To test PROGEN we reconstructed 60 different protein structures (or complexes) extracted from the BPDB. Comparison between PROGEN model and BPDB reference structures gave r.m.s.d. values for peptide main-chain atoms between 0.29 and 0.76 Å, with a grand average of 0.53 Å for all 60 models. The r.m.s.d. for all non-hydrogen atoms ranged between 1.44 and 1.93 Å for the 60 polypeptide models generated. In many instancess, the models constructed by PROGEN are comparable or superior to previous methods [2–5]. The correct assignments of *cis*- or *trans*-proline configurations were made in the test models and could be attributed to the OGP files for proline. PROGEN is fully automatic and only two sets of data (amino acid sequence and C^{α} atomic coordinates) are required; no additional structural information is required.

MATERIALS AND METHODS

Instrumentation and software

Modelling and computations were performed in our laboratory on Silicon Graphics IRIS 210GTX and 70G computer workstations (Silicon Graphics, Mountain View, CA) and a VAX750/VMS system (Digital Equipment Corp.). QUANTA 3.0 (POLYGEN Corp., Waltham, MA), with attached CHARMm21 program [10], was used for energy minimization (EM), structural analysis and graphic display on the IRIS workstations. The solvent accessibility algorithm, ACCESS [11], was run on the VAX750. PROGEN and all other FORTRAN software were entirely developed in our facility for use on the IRIS workstation.

The modelling strategy of PROGEN

The complete modelling procedure is carried out in two major phases: (i) initial generation of the peptide, and (ii) refinement techniques involving EM and dynamics simulations [12]. A detailed text explaining each of the steps and data inputs involved in the PROGEN algorithm is provided in the Appendix (summarized in the Appendix Table 1). The α-carbon coordinates for the protein structures used in this study were extracted from the atomic coordinates provided in the Brookhaven Protein Database (BPDB) [1]. The OGP database was derived from 19 BPDB protein structures using a FORTRAN program developed in our laboratory (the structures are listed in the Appendix Table 2). Correlations between 23 different bond angles, dihedral angles, and bond distances (shown in Fig. 1 and defined in the Appendix Table 3) and the α-carbon distances for each amino acid were measured using a FORTRAN program. These correlation data are presented in the Appendix Table 4.

The most difficult task in the generation of the structure was the placement of the main-chain atoms at their proper inter-residue positions, thereby defining the main-chain conformation. PROGEN solved this problem by rotating the atoms of an amino acid residue around its fixed α-carbon atom in the 3D space and calculated the deviations of the measured parameters from the predicted OGP values; this procedure identified the orientation with the minimum r.m.s.d. of all the inter-residue parameters. Details of this procedure are provided in the Appendix.

Reliability of PROGEN generated structures

In order to check the reliability of the PROGEN algorithm, we selected and modelled 60 known protein structures (and sub-units) from the BPDB. Using a FORTRAN program (MakeCApdbFile), the coordinates of only the α-carbon atoms were extracted from the original BPDB files and stored in separate ASCII files for all the 60 polypeptide structures. Using these Ca_pdb_Files the complete polypeptides were generated with the PROGEN algorithm as 'Final_Construct.pdb' files, containing all the atomic coordinates. These reconstructed structures were then compared with the original BPDB reference structures by several different procedures using QUANTA and other FORTRAN programs.

Deviations from the reference structure atomic positions

Using the 'Comparison' functions of QUANTA, the PROGEN-generated and BPDB structures were optimally aligned with respect to α-carbons, peptide main chains, and all atoms (individually for these three selections), followed by the recording of the r.m.s.d., along with the

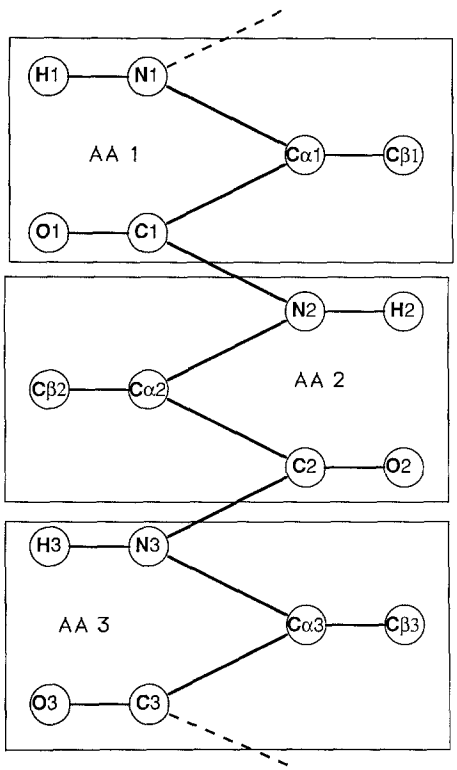

Fig. 1. The tripeptide segment of a polypeptide showing the names and numbers of the atoms in the residues as defined in this study. The distances, angles and dihedral angles of the intra- and inter-residue geometry parameters (as defined in Appendix Table 3) are described by these atom names and numbers.

individual r.m.s.d. of all the atoms. The individual r.m.s.d. were analyzed for each amino acid, as well as for the secondary structures of the polypeptides.

Deviations from the reference structure bond distances, bond angles and dihedral angles

The deviations of the various conformational and configurational parameters of the PROGEN-generated structures from the reference structures were examined. The correlations between the parameters of the constructed and the BPDB structures were examined. The φ–ψ dihedral angle plots for specific residues of the PROGEN and reference structures were also compared.

Solvent-accessible surfaces and volumes

The solvent-accessible surface areas of the PROGEN-generated and BPDB reference structures were calculated according to the ACCESS program of Lee and Richards [11] which uses a spherical probe of water (r = 1.4 Å) and calculates the contact area on the VDW surface of each atom that can be contacted by the probe (denoted as 'solvent-accessible surface' area). The VDW volumes of the polypeptide sets were measured using QUANTA and compared.

Introduction of statistical errors to the C^{α} coordinates

To determine if PROGEN could build accurate structures from unrefined α-carbon coordinates, we introduced random statistical errors to the refined coordinates for several test cases using a FORTRAN program. For each α-carbon atom in the data set, the program generated three random numbers betwee −1.0 and 1.0 and the product of these numbers, with a preset maximum deviation, was added to the X, Y and Z values of the atomic coordinate. A number of different sets of α-carbon coordinates were generated with increasing values of the maximum deviation. Complete structures were modelled from these 'error-containing' data sets and compared with the refined structures of the BPDB.

RESULTS AND DISCUSSION

The comparison of atomic deviations for PROGEN and BPDB reference structures

Table 1 shows the atomic deviations of the 60 PROGEN-generated structures and the BDPB reference structures when optimally aligned using QUANTA. There was little deviation observed when only the α-carbon atoms were compared (data not shown) because the model C^{α} positions were derived from the original reference structures; these atoms remained fixed to their original coordinates during the entire modelling process. The r.m.s.d. of the peptide main-chain atoms for the 60 models (Table 1) ranged between 0.29 to 0.76 Å giving a grand average value of 0.54 Å. The average atomic deviations of the main-chain atoms were between 0.14 and 0.44 Å with a grand average value of 0.28 Å for the entire series. The r.m.s.d. of all non-hydrogen atoms ranged between 1.51 to 1.93 Å and the average deviations were between 0.73 and 1.16 Å. Although no errors were introduced into the α-carbon coordinates in these experiments, we did examine PROGEN models containing random errors introduced into the original α-carbon coordinates (see below).

Other investigators have devised different approaches to the construction of all atom coordinates for proteins starting with only the α-carbon coordinates [2–5]. A comparison of the models created by PROGEN and other algorithms is presented in Table 2. Reid and Thornton [2,3] modelled several protein molecules using a dictionary of peptide backbone structures and statistical distributions for amino acid side-chain conformations. Their initial model of flavodoxin (3FXN), for example, had an atomic r.m.s.d. of 0.57 and 1.73 Å, for main chains and all atoms, respectively, when compared to the crystallographic coordinates. Correa [4] modelled 3FXN by the sequential addition of backbone amino acids with Pro, Gly and Ala (for all AA) followed by the generation of side chains, combined with molecular dynamics refinements at each step. This approach produced a model with atomic r.m.s.d. of 0.49 and 1.64 Å, for main chains and all atoms, respectively. Recently, Holm and Sander [5] generated the atomic coordinates of a number of polypeptides, including flavodoxin, from α-carbon atom traces using a database algorithm. The r.m.s.d. of their 3FXN model were reported to be 0.48 and 1.57 Å, for main chains and all atoms, respectively. The corresponding r.m.s.d. in our PROGEN model of 3FXN were 0.46 Å (main chain) and 1.71 Å (all atoms) and are comparable to the above-cited studies. Correa modelled α-lytic protease (2ALP) to obtain r.m.s.d. of 0.19 Å (main chain) and 1.24 Å (all atoms) [4]. For the model of troponin C (5TNC) r.m.s.d. of 0.41 Å (main chain) and 1.68 Å (all atoms) were attained. The corresponding r.m.s.d. obtained using PROGEN models were 0.35 and 1.54 Å for α-lytic protease (2ALPS in Table 2) and 0.36 and 1.72 Å for troponin C (5TNC in Table 2).

TABLE 1
COMPARISON BETWEEN THE PROGEN MODELS AND THE BPDB STRUCTURES

Structure	r.m.s.d. (Å)		Structure	r.m.s.d. (Å)	
	Main chain	All atoms		Main chain	All atoms
1REIA	0.48	1.91	3HHBB	0.32	1.62
1REIB	0.37	1.85	3P2PA	0.60	1.84
1TIMS	0.53	1.64	3RNTS	0.56	1.62
2ALPS	0.36	1.54	4FDLS	0.54	1.54
2CI2I	0.43	1.63	4FXNS	0.48	1.84
2CTSS	0.44	1.81	4GPD1	0.75	1.78
2FB4H	0.45	1.66	4GDP2	0.73	1.75
2FB4L	0.46	1.75	4GDP3	0.75	1.82
2HFLH	0.64	1.90	4GDP4	0.71	1.79
2HFLL	0.48	1.90	4HHBA	0.38	1.88
2HFLY	0.57	1.77	4HHBB	0.37	1.69
2IG2H	0.54	1.80	4HHBC	0.41	1.84
2IG2L	0.51	1.63	4HHBD	0.37	1.56
2RHES	0.44	1.44	4MDHA	0.44	1.67
2SECE	0.41	1.56	4MDHB	0.44	1.70
2SECI	0.42	1.60	4PTIS	0.45	1.84
2SNIE	0.35	1.58	4PTPS	0.43	1.75
2SNII	0.44	1.56	4XIAA	0.50	1.69
2SODB	0.57	1.53	4XIAB	0.55	1.76
2SODG	0.71	1.69	5CPAS	0.45	1.85
2SODO	0.70	1.48	5CYTS	0.39	1.73
2SODY	0.76	1.61	5MBNS	0.40	1.83
3ADKS	0.46	1.60	5PEPS	0.49	1.57
3CLNS	0.31	1.71	5TNCS	0.35	1.68
3FXNS	0.46	1.64	7ADHS	0.63	1.62
3GRSS	0.41	1.50	7TLNS	0.46	1.56
3HFMH	0.68	1.79	8DFRS	0.53	1.81
3HFML	0.58	1.89	9PAPS	0.49	1.92
3HFMY	0.64	1.85	9WGAA	0.54	1.58
3HHBA	0.28	1.90	9WGAB	0.57	1.65

Grand average values of the main-chain and all atom r.m.s.d. were 0.51 and 1.71 Å, respectively. The r.m.s.d. for main chain ranged between 0.28 and 0.76 Å. The r.m.s.d. for all atoms ranged between 1.44 and 1.92 Å.

The model of 2ALP by Correa involved refinements based on the prior knowledge of the presence of *cis*-proline and special consideration was required for this situation. In contrast, all of our results with PROGEN were obtained by applying a uniform automatic batch procedure for all the polypeptides examined in this study. No special considerations or prior knowledge of specific structural features for any given polypeptide were required. For example, in our model of 2ALP, the orientation of the *cis*-proline residue was correctly assigned because PROGEN makes the assignment based on the OGP file for *cis*- and *trans*-proline residues and the C^{α}–C^{α} distances contained therein (see below).

The use of a rotamer library from known structures may help model structures more closely to the crystallographically determined positions, but may not accurately represent their configurations in solution [13]. For side-chain atom placement PROGEN did not use a specific side-chain rotamer library and, consequently, the atomic deviations for all atoms were greater (about 0.3 Å) than those established in the studies by Holm and Sander [5] and Tuffery et al. [14] in which their side chain atoms were positioned using a rotamer library. Placement of amino acids by PROGEN was accomplished using the optimal geometry parameters along the main-chain and placement of the β-carbon atoms close to their optimal positions, without consideration of other side-chain atoms. After the initial coordinate files were created, the side-chain atoms were rotated along the α–β and β–γ bonds in order to place atoms in positions with minimum contacts. The EM and dynamics simulations placed side chains to energetically favorable positions. The positions of the side chains on the exposed surfaces of the molecule are probably highly influenced by the solvent and crystal-packing environment. Additional modelling studies may prove helpful in predicting side-chain configurations.

Energy minimization refinements

Table 3 presents a few examples of the calculated potential energy values for the models at different steps of the modelling refinement. In Step 1, the atomic deviations from the main chain atoms were fairly close to the final values in the initial construction of the full structure and the energy values were very high due to some unusual contacts of the side chain atoms. Because no energy terms were considered in the initial process, some unusual bond and angle situations resulted in unfavorable conformations along the chain. Some of these strains were relieved in Step 2 at which time the side chains were rotated for the best positional conformations. The EM procedure carried out in Step 3 improved the polypeptide structure resulting in sharp reductions

TABLE 2
COMPARISON OF PROGEN WITH OTHER C^{α} MODELLING METHODS

Protein	r.m.s.d. (Å) from BPDB structure		
	Method[a]	Main chain	All atoms
3FXNS	1	0.57	1.73
	2	0.49	1.64
	3	0.48	1.57
	4	0.46	1.71
2ALPS	2	0.19	1.24
	4	0.36	1.54
5TNCS	2	0.41	1.68
	4	0.36	1.52
1TIMS	3	0.59	–
	4	0.53	1.64
2CTSS	3	0.45	–
	4	0.43	1.81
5CPAS	3	0.48	–
	4	0.44	1.85

[a] Modelling method used: 1, Reid and Thornton [3]; 2, Correa [4]; 3, Holm and Sander [5]; 4, PROGEN (this report).

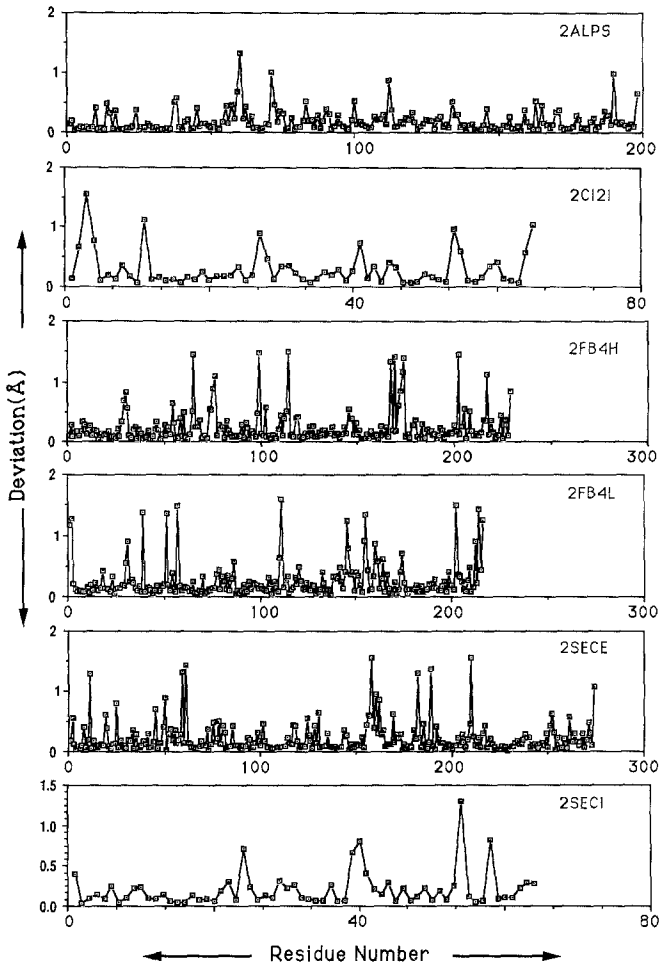

Fig. 2. The r.m.s.d. (Å) of the main-chain atoms of the polypeptide structures generated by PROGEN from the BPDB atomic coordinates along the chain starting from the amino-terminus. The deviations were measured using the 'Comparison' function of QUANTA after the optimal alignment of the constructed structures with the known structures. Data for 6 polypeptides, 2ALPS, 2CI2I, 2FB4H, 2FB4L, 2SECE and 2SECI are shown.

of the potential energy values. Despite EM procedures, the deviations in the atomic coordinates improved only slightly (data not shown). Using dynamics simulations and subsequent EM there were few changes in the atomic r.m.s.d. The average times required at each step per residue are shown at the bottom of Table 3. The molecular dynamics simulations require the most time and were not very effective in reducing the atomic deviations. The use of other algorithms for dynamic heating and annealing, such as XPLOR [15], may prove more useful in this regard.

The distribution of r.m.s.d. along the polypeptide chain

Figure 2 shows the plot of average atomic deviations of the main chain of individual amino acids along the polypeptide chains for six polypeptides. The deviations along the chain are not

TABLE 3
POTENTIAL ENERGIES (KCAL/MOL) AT VARIOUS STEPS OF REFINEMENT FOR DIFFERENT PROTEIN STRUCTURES

Name	Step 1	Step 2	Step 3	Step 4
2ALPS	*******	61015	−5628	−6068
2CI2I	1165533	484035	−1940	−2515
2FB4H	*******	*******	−5348	−6668
2FB4L	*******	518465	−6026	−6870
2IG2H	*******	*******	−5992	−6846
2IG2L	*******	74465	−5871	−6675
2SECE	*******	98539	−7912	−8582
2SECI	*******	51203	−1652	−2018
2SNIE	*******	76621	−7966	−8622
2SNII	*******	696281	−1961	−2330
2SODB	*******	50376	−4393	−5048
2SODG	*******	409735	−4155	−4791
2SODO	*******	51246	−4157	−4850
2SODY	*******	424276	−4118	−4935
3CLNS	*******	4359	−4365	−5146
3FXNS	*******	32350	−4596	−4974
3GRSS	*******	*******	−13615	−15888
3HFMH	*******	*******	−5066	−5924
3HFML	*******	*******	−5997	−6843
3HFMY	*******	30165	−3786	−4018
3HHBA	*******	4418	−4280	−4881
3HHBB	*******	*******	−4426	−4931
4FDLS	*******	11570	−3060	−3513
4FXNS	*******	49500	−4427	−5444
4GPD1	*******	3775225	−8296	−10024
4GPD2	*******	6281203	−8510	−10249
4GPD3	*******	57538	−8619	−10246
4GPD4	*******	1298925	−8674	−10396
4HHBA	*******	*******	−3998	−4572
4HHBB	5660015	16393	−4044	−4568
4HHBC	*******	55665	−3918	−4586
4HHBD	*******	51198	−3979	−4671
4MDHA	*******	64242	−10129	−11789
4MDHB	*******	1135008	−10090	−11523
4PTPS	*******	39706	−6360	−7034
5CYTS	*******	3316	−3284	−3767
5MBNS	*******	14981	−5067	−5846
5TNCS	*******	280129	−5365	−6300
7ADHS	*******	*******	−9991	−11733
7TLNS	*******	6438349	−9464	−10608

Step 1: Generation of the 'Initial_Construct.pdb' from 'Ca_pdb_File' using PROGEN (lines 1–3 in Appendix Table 1).
Step 2: Rotation of the side chains of the polypeptide using QUANTA (lines 4–12 in Appendix Table 1).
Step 3: Energy minization using CHARMm (lines 13–16 in Appendix Table 1).
Step 4: Dynamic heating and cooling using CHARMm (lines 17–19 in Appendix Table 1).
Average computer time (s/residue) on IRIS 210 GTX for steps 1–4 was 16.0, 9.5, 6.7 and 28.4, respectively.
*******Value is > 9999999.

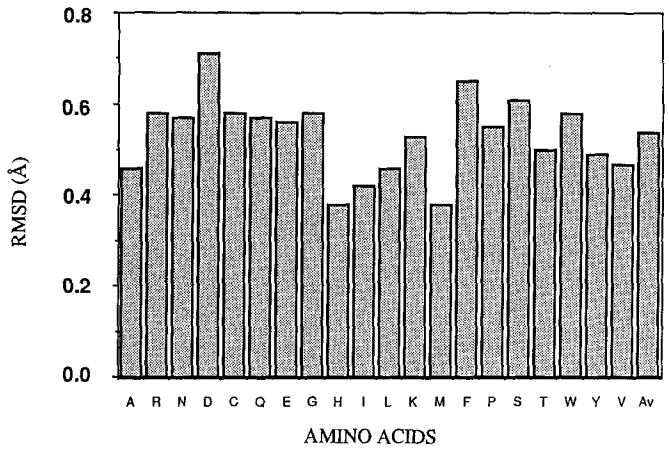

Fig. 3. r.m.s.d. (Å) of the main-chain atoms of the polypeptide structures generated by PROGEN from the crystallographically determined atomic coordinates with respect to specific amino acids. Deviations were calculated for 60 polypeptides (from Table 1) containing more than 16 000 residues. The bar at the far right (av) shows the grand average of the deviations of all the amino acids.

randomly distributed. There are, in fact, some regions of higher deviations for all the polypeptide chains and these regions are clearly visible, especially in the cases of certain smaller polypeptides (2CI2 and 2SECI). For longer polypeptides the peak deviations are observed only when plotted in expanded scales. These localized peaks suggest that the mismatches are cooperative phenomena. Perhaps an imperfect orientation of one residue perturbs the orientations of its neighbours. The identification of these regions by some other means than comparing with the known structure (which would not be available in the real predictive application of this method) would provide a guide for further refinement of the modelled structure. Conformational search algorithms, such as CONGEN [16], may provide additional low-energy structures and thereby provide further refinement of these regions.

Dependence of r.m.s.d. on the amino acids

We examined the modelled structures to identify the specific amino acids, if any, which may have been associated with areas of significant deviations from the known structures. Figure 3 shows the deviations of the main-chain atoms of all the 20 amino acids combined with the average of all the residues. Data were collected from more than 16 000 amino acids contained in the 60 polypeptides. The grand r.m.s.d. for all the amino acids was 0.54. The r.m.s.d. of all the amino acids fall within 30% of the grand averages. Methionine (M) and histidine (H) showed minimum r.m.s.d. (0.38) and aspartate (D) showed maximum r.m.s.d. (0.71).

Dependence of r.m.s.d. on the secondary structures

Table 4 presents three sets of the typical values for the average and r.m.s.d. according to the secondary structures. The first set shows the corresponding values for one (2IG2L) of the 60 polypeptides examined; the second set is for one (alanine) of the 20 amino acids from all the polypeptides examined; the third set is the grand average of all the amino acids in all the polypeptides studied. From the third set, which considers all the amino acids in all the polypeptides, it

may be inferred that the α-helical, β-strand and extended regions showed lower deviations compared to folded β-turns and undefined secondary structural regions. Almost the same pattern is observed when the amino acids are examined individually (second set in Table 4).

In Table 5 the secondary structures of three pairs of constructed and BPDB reference polypeptide molecules are shown. These data indicate that the constructed structures very closely resemble those of the reference structures in terms of secondary structure and intra-molecular hydrogen bonding. Graphical representation of the secondary structures and hydrogen bonds of the constructed and reference molecules of 2ALP were found to be superimposable (data not shown).

Effect of cis-prolines on the structure predictions

When the atomic r.m.s.d. of the main-chain atoms were compared, proline had a value very close to the grand average of all the amino acids (Fig. 3). But when the *cis*- and *trans*-configurations of proline were segregated (Table 6), it was found that the r.m.s.d. was higher at *cis*-prolines (0.71 Å) than that at *trans*-prolines (0.54 Å). A similar problem was encountered by Correa while building the structure of 2ALP and special treatments were required for modelling near the *cis*-proline residues. In our procedure, all the amino acids were oriented using the same algorithm,

TABLE 4
RMSD OF THE MAIN-CHAIN ATOMS ACCORDING TO THE SECONDARY STRUCTURES

Data from	Secondary structure	No. of amino acids	%	r.m.s.d. (Å)
Polypeptide 2IG2L	α-Helix	16	7.4	0.35
	β-Strand	95	44.0	0.39
	Extend	24	11.0	0.42
	Folded	2	0.9	0.23
	β-Turn	31	14.4	0.73
	Random	48	22.2	0.61
	Total	216	100.0	0.51
All alanine	α-Helix	311	44.5	0.29
	β-Strand	89	12.7	0.39
	Extend	63	9.0	0.47
	Folded	20	2.9	0.58
	β-Turn	77	11.0	0.55
	Random	139	19.9	0.69
	Total	699	100.0	0.46
All amino acids	α-Helix	2239	27.9	0.33
	β-Strand	1530	19.1	0.43
	Extend	787	9.8	0.47
	Folded	208	2.6	0.74
	β-Turn	1162	14.5	0.66
	Random	2103	26.2	0.69
	Total	8028	100.0	0.54

but different OGP files were used, based on the particular amino acid involved, and their α-carbon distance parameters. Thus, all the proline residues received the same OGP values for the same C^{α} environments, viz., the distance between the C^{α} atom of a *cis*-proline and the C^{α} atom of the preceding amino acid (towards N-terminus) is 2.7 Å, whereas it is 3.8 Å for the *trans*-configuration. Therefore, in the OGP files all the *cis*-prolines are modelled as *cis*, and the *trans*-prolines are modelled as *trans*.

Correlation of the parameters between the PROGEN and reference structures

The conformational and configurational parameters of the constructed and reference structures were correlated to check the validity of such structural predictions. Figure 4 presents the correlation diagrams for the important dihedral angles: φ (Panel A) and ψ (Panel B) of 3GRSS, the largest single polypeptide chain modelled by PROGEN (461 residues). The extent of correlation was very good (correlation coefficient values 0.83 and 0.85 for φ and ψ, respectively). Panels C and D of Fig. 4 show the correlation diagrams for the angles aSFC (0.73) and aFSC (0.78). Although these angles are not bond angles, they help orient the amino acids to their correct alignments due to their consistent values within a narrow limit (around 13° in the case of aFSC and 21° in the case of aSFC). Two other angles of a similar nature, involving the main-chain nitrogen atoms, also showed very narrow ranges of values (i.e., 15° for aFSN and 9° for aSFN). These angles showed a remarkable dependence on the α-carbon atom distance parameters for some amino acids (Appendix, Table 4). The mean values of the other distance and angle parameters of the constructed structures were very close to their OGP values (data not shown).

Comparison of solvent-accessible surfaces (SAS)

The SAS areas of some of the PROGEN-generated structures were calculated and compared with those of the reference structures. Figure 5 presents a plot of the SAS areas in terms of individual residues for 2ALPS with a surface probe the size of a water molecule (r = 1.4 Å). The PROGEN structure was found to have the same regions of SAS as on those calculated for the reference structures, with slight quantitative variations. Table 7 presents the total SAS, VDW

TABLE 5
COMPARISON OF THE SECONDARY STRUCTURES AND MAIN-CHAIN HYDROGEN BONDS IN PROGEN-GENERATED AND BPDB STRUCTURES

	2ALPS		3GRSS		5MBNS	
Secondary structure	PROGEN (%)	BPDB (%)	PROGEN (%)	BPDB (%)	PROGEN (%)	BPDB (%)
α-Helix	3.6	3.5	28.2	29.5	70.4	77.0
β-Strand	47.7	45.0	15.6	17.1	0.0	0.0
Extend	9.1	7.1	14.1	13.7	0.7	0.0
Folded	1.5	3.0	1.7	1.7	0.7	0.7
β-Turn	18.8	17.7	16.9	14.5	15.8	9.0
Undefined	19.3	23.6	23.4	23.4	12.5	13.3
No. of H bonds	115	119	281	280	115	119

areas and VDW volumes for the comparison of PROGEN and reference structures for three molecules: 2ALPS, 4FXNS and 5MBNS. For 2ALPS the PROGEN structure is slightly more 'open' than the reference structures, as it has a slightly larger SAS, VDW areas and VDW volume. Comparison of the calculated surface areas for PROGEN and reference structures of 4FXNS are closely matched (within 20–160 Å^2); the VDW volumes for the two 4FXNS structures are nearly identical.

Effect of random statistical errors in the C^{α} coordinates

Table 8 presents the r.m.s.d. of the structures of the α-chain of hemoglobin (3HHB), modelled from different sets of α-carbon coordinates with introduced statistical errors in the C^{α} positions (maximum errors of 0.0–1.8 Å). Even with introduced errors greater than 1.8 Å a complete structure could be built, although the r.m.s.d. values as compared to the BPDB structure increased with increasing error introduction. Final models were obtained after refinement in which the α-carbon atoms were allowed to move during EM using SD, CG and ABNR (10 steps

TABLE 6
AVERAGE AND RMSD (Å) OF THE MAIN-CHAIN ATOMS AT PROLINE RESIDUES

Secondary structure	No. of AA	%	RMSD
All prolines			
α-Helix	43	12.4	0.32
β-Strand	42	12.1	0.48
Extend	38	10.9	0.58
Folded	11	3.2	0.90
β-Turn	98	28.2	0.60
Undefined	116	33.2	0.54
Total	348	100.0	0.55
cis-Prolines only			
α-Helix	0	0.0	–
β-Strand	8	37.2	0.72
Extend	2	9.3	0.48
Folded	1	4.7	1.62
β-Turn	2	9.3	0.70
Undefined	8	39.5	0.55
Total	21	100.0	0.71
trans-Prolines only			
α-Helix	43	13.1	0.32
β-Strand	34	10.4	0.40
Extend	36	11.0	0.58
Folded	10	3.1	0.79
β-Turn	96	29.4	0.60
Undefined	108	33.0	0.54
Total	327	100.0	0.54

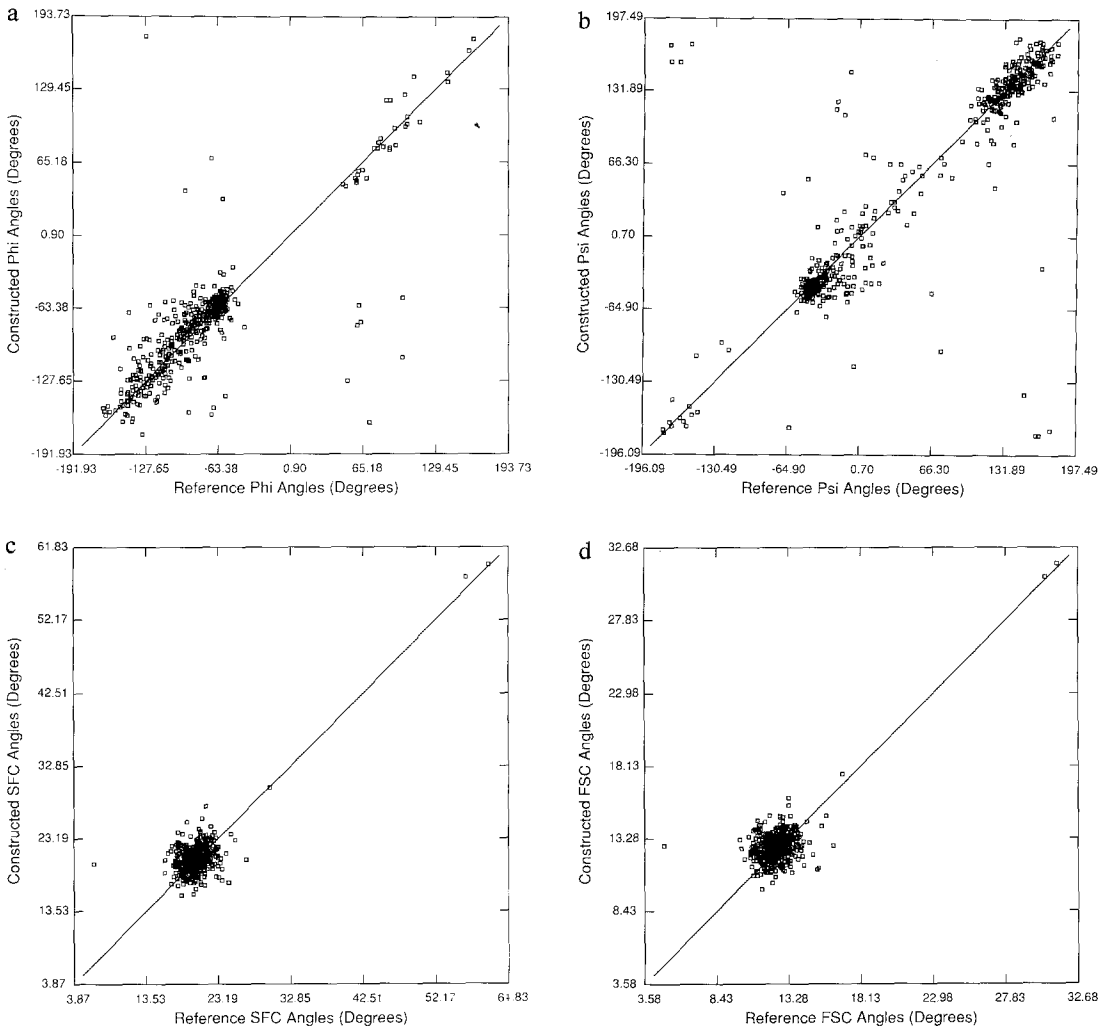

Fig. 4. Correlation diagrams of some angular parameters of the PROGEN and BPDB reference structure of the polypeptide 3GRSS. Panel a is for φ angles; panel b is for ψ angles; panel c is for SFC angles; and panel d is for FSC angles.

each) with dihedral constraints. After this EM, the r.m.s.d. of the main chain including α-carbons decreased; especially with higher maximum errors (0.6 Å or greater), when compared to the BPDB structures.

Models based on BPDB C^{α} structures

Table 9 presents PROGEN models constructed from unrefined BPDB C^{α} atomic coordinates as compared to models constructed from completed and refined BPDB structures. The C^{α} atomic coordinates of lysozyme (1LZH), containing 2 polypeptides (A and B), were obtained from unrefined crystallographic data; PROGEN models of polypeptides 1LZHA and 1LZHB were then compared with the refined structures of lysozyme 2HFL (complexed with its antibody,

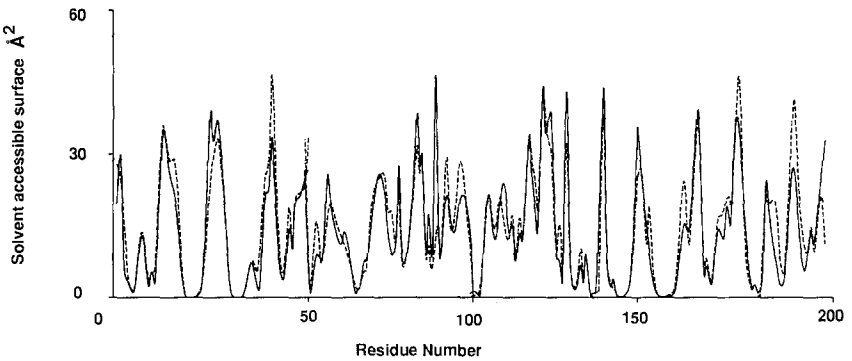

Fig. 5. Comparison of the solvent-accessible surfaces of the individual residues (from N-terminus) of the PROGEN constructed (dashed line) and BPDB reference (solid line) structures of 2ALPS.

R-value 0.254) and 1LYM (R-value 0.260). The r.m.s.d. with respect to C^{α}, main chain, and all atoms, were comparable to the data obtained using a C^{α} data set containing an introduced error of 1.0 Å or greater. Similar r.m.s.d. were also obtained when the PROGEN structure generated from the α-carbon coordinates of the immunoglobulin λ-chain dimer (1MCG, C^{α} coordinates only, R-value 0.371) was compared to the refined structure (2MCG, complete coordinates, R-value 0.187). These PROGEN-generated structures are in agreement with the complete protein structures.

CONCLUSION

We have developed a fully automated modelling algorithm, PROGEN, for the generation of the complete atomic coordinates of a protein polypeptide starting with only the knowledge of the atomic coordinates of its α-carbon atoms and the complete amino-acid sequence. The first part of PROGEN analyzed a number of intra-peptide and inter-peptide configurational and conformational parameters for amino-acid residues found in 19 empirically solved protein structures of the BPDB. These parameters were analyzed and optimal geometry parameter files were established for each amino acid. This use of the BPDB differs from that of other modelling algorithms,

TABLE 7
COMPARISON OF THE SOLVENT-ACCESSIBLE SURFACE, VAN DER WAALS AREAS AND THE VAN DER WAALS VOLUMES OF PROGEN-GENERATED AND BPDB STRUCTURES

Structure	2ALPS		4FXNS		5MBNS	
	PROGEN	BPDB	PROGEN	BPDB	PROGEN	BPDB
Solvent-accessible surface area (Å^2)	8 620	8 348	6 891	7 157	8 491	8 601
van der Waals area (Å^2)	2 767	2 652	2 183	2 206	2 762	2 746
van der Waals volume (Å^3)	19 681	19 553	15 509	15 531	17 526	17 897

The solvent-accessible surface areas were calculated using the program ACCESS (calculated as the area accessible to a 1.4 Å sphere on the VDW surface); the total van der Waals areas and volumes were obtained calculated using QUANTA.

TABLE 8
INTRODUCTION OF STATISTICAL ERRORS TO THE C^{α} COORDINATES AND SUBSEQUENT PROGEN MODELS

Structure	Maximum		RMSD (Å)	
	Error (Å)	C^{α} only	Main chain	All atoms
3hhb0	0.0	0.11	0.15	1.62
3hhb1	0.2	0.20	0.45	1.75
3hhb2	0.2	0.19	0.47	1.81
3hhb3	0.4	0.20	0.50	1.80
3hhb4	0.4	0.22	0.49	1.68
3hhb5	0.6	0.30	0.75	1.93
3hhb6	0.6	0.28	0.66	1.89
3hhb7	0.8	0.36	0.82	1.83
3hhb8	0.8	0.36	0.78	1.82
3hhb9	1.0	0.43	0.86	1.82
3hhb10	1.0	0.43	0.78	1.88
3hhb11	1.4	0.62	1.17	2.12
3hhb12	1.4	0.62	1.03	2.06
3hhb13	1.6	0.66	1.23	2.18
3hhb14	1.6	0.69	1.27	2.21
3hhb15	1.8	0.77	1.30	2.42
3hhb16	1.8	0.83	1.28	2.30

in that the statistical analysis of the geometric information from the BPDB is extracted and stored as an OGP file in the form of mathematical expressions for later use. We have deduced the geometric relationships for 23 distances and angles of polypeptide backbones, and their correlations with the C^{α} atom distances. Purisima and Scheraga [17] derived the inter-relations of φ and ψ dihedral angles with the C^{α} distances using an analytical approach, but our derivation of these parameters was based on known structures in the BPDB, and consequently, we found slightly different OGP values for various amino acids; this aspect may have aided in more accurate model predictions.

TABLE 9
COMPARISON OF PROGEN MODELS DERIVED FROM BPDB UNREFINED C^{α} COORDINATES WITH COMPLETED STRUCTURES

Structures		RMSD (Å)		
Unrefined	Completed	α-Carbon	Main chain	All atoms
1LZHA	2HFLY	0.46	0.78	1.79
1LZHB	2HFLY	0.46	0.74	2.03
1LZHA	1LYMA	0.54	0.87	1.93
1LZHB	1LYMB	0.60	0.91	2.12
1MCG1	2MCG1	0.67	0.89	1.88
1MCG2	2MCG2	0.68	0.93	1.65

During the generation of the polypeptide main chain, the side-chain atoms remained fixed relative to their respective main-chain atoms. After the initial orientation of the main-chain atoms, the side-chain atoms were placed in positions with minimum VDW contacts using QUANTA. During the EM procedures, the side-chain atoms, as well as the main-chain atoms, moved to positions which yielded local energy minima. We did not use any special algorithm for the placement of side-chains and the r.m.s.d. of the side-chain atoms in PROGEN-modelled structures are very close to those models developed by previous investigators using more robust EM techniques [3,4].

PROGEN is fully automatic and only two sets of inputs are to be provided. One set is in the UNIX environment, while the other set is in the QUANTA environment. PROGEN does not require any special structural information about the polypeptide under construction. We tested PROGEN on 60 polypeptides. This procedure will be very helpful in building proteins for which only the α-carbon atom coordinates are available. It may also be useful for the refinement of the preliminary structures obtained in the early stages of X-ray crystallographic or NMR determinations. An extension of this method may lead to the development of algorithms for the database-derived prediction of complete protein structures from the coordinates of fewer atoms than all the α-carbon atoms. PROGEN may also be useful to molecular biologists for the modelling specific amino-acid replacement or deletion mutations in known structures. We are currently testing PROGEN for the modelling of protein superfamilies that contain semi-homologous members having minor structural deletions or replacements. PROGEN is available from the authors upon written request (academic or other license agreement).

Editors' note

A paper [Levitt, M., J. Mol. Biol., 226 (1992) 507] regarding a similar approach has recently been brought to the Editors' attention.

ACKNOWLEDGEMENTS

This work was supported by grants from the National Institutes of Health/National Institute of Drug Abuse (DA-07240) and General Medical Science (GM-46535).

REFERENCES

1 Bernstein, F.C., Koetzle, T.F., Williams, E.J.B., Meyer Jr., E.F., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
2 Reid, L.S. and Thornton, J.M., In Protein Structure, Folding, and Design (UCLA Symposium NS69), 1987, pp. 93–103.
3 Reid, L.S. and Thornton, J.M., Proteins, 5 (1989) 170.
4 Correa, P.E., Proteins, 7 (1990) 366.
5 Holm, L. and Sander, C., J. Mol. Biol., 218 (1991) 183.
6 Mandal, C., Shirley, F., Anchin, J.A., Mandal, C. and Linthicum, D.S., Hybridoma, 10 (1991) 459.
7 Walter, D.E., Orthoefer, F.T. and DuBois, G.E. (Eds.) Sweeteners: Discovery, Molecular Design and Chemoreception (ACS Symposia, Series 450), American Chemical Society, Washington DC, 1991.
8 de Vos, A.M., Hatada, M., van der Wel, H., Krabbendam, H., Peerdeman, A. and Kim, S.-H., Proc. Natl. Acad. Sci. USA, 82 (1985) 1406.
9 Ogata, C., Hatada, M., Tomlinson, G., Shin, W.C. and Kim, S.-H., Nature, 328 (1987) 739.

10 Brooks, B.R., Bruccoleri, R.E., Olafson, H.D., States, D.J., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4 (1983) 187.
11 Lee, B. and Richards, F.M., J. Mol. Biol., 55 (1971) 379.
12 McCammon, J.A., Gelin, B.R. and Karplus, M., Nature, 267 (1977) 585.
13 Ponder, J.W. and Richards, F.M., J. Mol. Biol., 193 (1987) 775.
14 Tuffery, P., Etchebest, C., Hazout, S. and Lavery, R.J., Biomol. Struct. Dynam., 8 (1991) 1267.
15 Brunger, A.T., Krukowski, A. and Erickson, J.W., Acta Crystallogr., 46 (1991) 1267.
16 Bruccoleri, R.E. and Karplus, M., Biopolymers, 26 (1987) 137.
17 Purisma, E.O. and Scheraga, H.A., Biopolymers, 23 (1984) 1207.

APPENDIX

PROGEN algorithm

Appendix Table 1 lists the individual steps in the PROGEN-modelling procedure. The steps in the first column of the input list, numbered 1 through 9, are given in the UNIX environment, in descending order, either individually or as a batch. The remainder of the inputs is given in the QUANTA environment (POLYGEN Corporation, Waltham, MA). These files contain the macro-commands for the step-wise generation of the polypeptide structure and complete the computational functions. PROGEN runs a FORTRAN program which takes the ASCII file containing α-carbon coordinates named 'Ca_pdb_File' (No. 2 in Appendix Table 1) and generates the initial sets of atomic coordinates of all the atoms of the polypeptide; these are stored as the output file named 'Initial_Construct.pdb' (No. 3 in Appendix Table 1).

PROGEN utilizes the BPDB information on various intra-peptide and inter-peptide structural parameters, such as bond distances, bond angles, and dihedral angles relevant to the main-chain conformations of known proteins to derive a set of OGP files and correlations with α-carbon distances. The 19 proteins selected for the extraction of this structural information are presented in Appendix Table 2, along with their BPDB code names, constituent subunits, crystallographic resolution (Å) and structural refinement factors (R-values). These proteins were chosen for the study because they represented a spectrum of secondary structures and range of atomic resolutions. For the known 19 protein structures (or complexes) examined, 23 different distance and angular parameters (defined in Appendix Table 3) were measured using a FORTRAN program (MeasureBDA.e) developed in our laboratory to establish the optimal geometry parameter database.

Generation of the optimal geometry parameter (OGP) database

The OGP files are compiled and sorted as two sets; one set for amino acids and the other for polypeptides. The program, StatisticsBDA.e, is used for the statistical analysis of these data in an attempt to define any correlation between the α-carbon atom distances (No. 1 in Appendix Table 3) and other conformational parameters (e.g., various bonds angles as shown in Fig. 1). Linear regression analyses are performed for the values of the distance and angle parameters (No. 2–19 in Appendix Table 3) taking each of the three α-carbon distances (dAA12, dAA13, and dAA23) as independent variables. The angle $C^{\alpha}1$–$C^{\alpha}2$–$C^{\alpha}3$ (Fig. 1) is also used as an independent variable for the linear regression analysis, but this produces the same results as the distance (dAA13), between the first and the third α-carbon atom. For the dihedral angles (Nos. 20–23 in Appendix Table 3), the deviations from mean values are used for the linear regression analysis; this is

required because of the fact that the dihedral angles are measured in circular scales between −180 and +180° and the two extremes meet at the same point. As a consequence, if two points are situated at −179 and +179, they are only 2° apart, contrary to their angular distance of 358°. Their mean value should be either +180 or −180 (as both are the same) rather than their algebraic mean of 0°. In order to deal with such situations, we developed special algorithms to calculate the mean values (as well as their deviations from the mean) for dihedral angles.

Because these OGP are found to be well correlated (> 0.5) with the one of the α-carbon distances (i.e., dAA12, dAA23 or dAA13), the program was modified to identify the residues (one each for 20 amino acids) and the parameters (intra-residue only) of which gave the least deviation

APPENDIX TABLE 1
LIST OF DATA INPUTS REQUIRED FOR PROGEN

No.	Commands	Time[a]	Description
1	PROGEN	15.9	Runs a FORTRAN program which takes α-carbon atom coordinates as input and generates the initial file containing the coordinates of all the non-hydrogen atoms of the polypeptide.
2	Ca_pdb_File		The input file: contains α-carbon atom coordinates (in PDB format).
3	Initial_Construct.pdb		The output file: contains full atomic atomic coordinates (in PDB format).
4	Make_Bond_Break_File	0.04	Runs a FORTRAN program which generates a file containing commands to QUANTA for breaking irrelevant bonds between sulfur atoms and other atoms.
5	Initial_Construct.pdb		The input file: Same file as in No. 3.
6	Break_Unwanted_Bonds.rec		The output file: contains macrocommand strings to QUANTA.
7	Make_Quanta_Spin_File	0.02	Runs a FORTRAN program which generates a file containing commands to QUANTA for spinning the side-chains to generate conformations with least contacts.
8	Initial_Construct.pdb		The input file: Same file as in No. 3.
9	Spin_SideChains.rec		The output file: contains macrocommand strings to QUANTA.
10	@Read_Initial_Construct.rec	0.05	Macro commands for making the initial msf.
11	@Break_Unwanted_Bonds.rec	1.06	Inputs the file generated in No. 6.
12	@Spin-SideChains.rec	8.29	Inputs the file generated in No. 9.
13	@Add_Polar_Hydrogens.rec	0.72	Adds polar hydrogen atoms using CHARMm.
14	@Change_Name.rec	0.04	Changes the name of the msf.
15	@Apply_Atom_Constraints.rec	0.06	Fixes the atomic coordinates of the α-carbon atoms
16	@Minimize_Energy.rec	5.90	Minimizes energy using CHARMm.
17	@Dynamics_Heat_Cool.rec	22.2	Runs dynamic calculations.
18	@Keep_Intermediate.rec	0.04	Keeps the intermediate msf file.
19	@Minimize_Energy.rec	6.19	Minimizes energy using CHARMm.
20	@Write_Final_Construct.rec	0.04	Writes the coordinates in the PDB format.

The command lines 1 through 9 work in UNIX environment and those from 10 through 20 work in QUANTA (3.0) environment. The inputs may be given individually or as batches.

[a] Total computer time in seconds per residue on an IRIS 210 GTX workstation.

from the mean values of all the residues for a given amino acid. The atomic positions of all the non-hydrogen atoms of the OGP of each amino acid, with respect to their α-carbon atoms in a particular orientation, are stored in 20 different files named after each amino acid (laa*.XYZ). The atomic positions are then used by PROGEN for building the initial polypeptide. At this point in the procedure, the intra-residue atomic distances and angles of the amino-acid residues are set to their mean values.

The statistical analyzer program, StatisticsBDA.e, developed by us, also calculated the deviations of the individual parameters from the mean values. In order to predict the conformational and configurational parameters from the α-carbon atom distances, the slopes, intercepts, correlation coefficients, and mean values were calculated using dAA12, dAA13 and dAA23 as the X-variable (one at a time) and the remainder of the parameters were used individually as the Y-variable. The set of X- and Y-variables which gave good correlation coefficients for the majority of amino acids were appended in the OGP amino acid files (laa*.XYZ) along with the

APPENDIX TABLE 2
BPDB PROTEINS USED FOR THE DERIVATION OF THE OGP DATABASE

No.	BPDB code	Subunit code	Protein structure	Resolution (Å)	R-value
1	2CI2	(2CI2I)	Chymotrypsin inhibitor2	2.0	0.198
2	2FB4	(2FB4H & 2FB4L)	Immunoglobulin Fab	1.9	0.189
3	2IG2	(2IG2H & 2IG2L)	Immunoglobulin G1	3.0	0.207
4	2SEC	(2SECE & 2SECI)	Subtilisin Carlsberg Complex with genetically-engineered N-acetyl Eglin-C	1.8	0.136
5	2SNI	(2SNIE & 2SNII)	Subtilisin novo complex with chymotrypsin inhibitor	2.1	0.154
6	3CLN	(3CLNS)	Calmodulin	2.2	0.175
7	3GRS	(3GRSS)	Glutathione reductase oxidized form(E)	1.5	0.186
8	3HFM	(3HFMH, 3HFML & 3HFMY)	IgG1 Fab fragment and lysozyme	3.0	0.246
9	3HHB	(3HHBA & 3HHBB)	Hemoglobin (deoxy)	1.7	0.200
10	4FD1	(4FD1S)	Ferredixin	1.9	0.212
11	4FXN	(4FXNS)	Flavodoxin (semiquinone form)	1.8	0.200
12	4GPD	(4GPD1, 4GPD2, 4GPD3 & 4GPD)	Apo-D-Glyceraldehyde-3-phosphate dehydrogenase	2.8	0.218
13	4MDH	(4MDHA & 4MDHB)	Cytoplasmic malate dehydrogenase	2.5	0.218
14	4PTP	(4PTPS)	β-Trypsin, diisopropyl-phosphoryl inhibited	1.34	0.171
15	5CYT	(4CYTS)	Cytochrome c (reduced)	1.5	0.159
16	5MBN	(5MBNS)	Myoglobin (deoxy)	2.0	0.179
17	5TNC	(5TNCS)	Troponin C	2.0	0.155
18	7ADH	(7ADHS)	Iso-nicotinimidylated liver alcohol dehydrogenase	3.2	0.290
19	7TLN	(7TLNS)	Thermolysin complex with CH_2CO(N-OH)LEU-OCH_3	2.3	0.170

APPENDIX TABLE 3
DISTANCE (d), ANGLE (a), AND DIHEDRAL (h) GEOMETRY PARAMETERS ANALYZED BY PROGEN

No.	Symbol	Atoms
1	dAA12	$C^{\alpha}1$–$C^{\alpha}2$
	dAA13	$C^{\alpha}1$–$C^{\alpha}3$
	dAA23	$C^{\alpha}2$–$C^{\alpha}3$
2	dCN(1)	C1–N2
	dCN(2)	C2–N3
3	dAN(1)	$C^{\alpha}1$–N2
	dAN(2)	$C^{\alpha}2$–N3
4	dCA(1)	C1–$C^{\alpha}2$
	dCA(2)	C2–$C^{\alpha}3$
5	dON(1)	O1–N2
	dON(2)	O2–N3
6	dNA(1)	N1–$C^{\alpha}1$
	dNA(2)	N2–$C^{\alpha}2$
	dNA(3)	N3–$C^{\alpha}3$
7	dAC(1)	$C^{\alpha}1$–C1
	dAC(2)	$C^{\alpha}2$–C2
	dAC(3)	$C^{\alpha}3$–C3
8	dCO(1)	C1–O1
	dCO(2)	C2–O2
	dCO(3)	C3–O3
9	dNC(1)	N1–C1
	dNC(2)	N2–C2
	dNC(3)	N3–C3
10	dAO(1)	$C^{\alpha}1$–O1
	dAO(2)	$C^{\alpha}2$–O2
	dAO(3)	$C^{\alpha}3$–O3
11	aACN(1)	$C^{\alpha}1$–C1–N1
	aACN(2)	$C^{\alpha}2$–C2–N3
12	aOCN(1)	O1–C1–N2
	aOCN(2)	O2–C2–N3
13	aCNA(1)	C1–N2–$C^{\alpha}2$
	aCNA(2)	C2–N2–$C^{\alpha}3$
14	aNAC(1)	N1–$C^{\alpha}1$–C1
	aNAC(2)	N2–$C^{\alpha}2$–C2
	aNAC(3)	N3–$C^{\alpha}3$–C3
15	aACO(1)	$C^{\alpha}1$–C1–O1
	aACO(2)	$C^{\alpha}2$–C2–O2
	aACO(3)	$C^{\alpha}3$–C3–O3
16	aFSN(1)	$C^{\alpha}1$–$C^{\alpha}2$–N2
	aFSN(2)	C^{α}–$C^{\alpha}3$–N3
17	aFSC(1)	$C^{\alpha}1$–$C^{\alpha}2$–C1
	aFSC(2)	$C^{\alpha}2$–$C^{\alpha}3$–C2
18	aSFN(1)	$C^{\alpha}2$–$C^{\alpha}1$–N2
	aSFN(2)	$C^{\alpha}3$–$C^{\alpha}2$–N3
19	aSFC(1)	$C^{\alpha}2$–$C^{\alpha}1$–C1
	aSFC(1)	$C^{\alpha}3$–$C^{\alpha}2$–C2
20	hPHI(1)	C1–N2–$C^{\alpha}2$–C2
	hPHI(2)	C2–N3–$C^{\alpha}3$–C3
21	hPSI(1)	N1–$C^{\alpha}1$–C1–N2
	hPSI(1)	N2–$C^{\alpha}2$–C2–N3
22	hNACO(1)	N1–$C^{\alpha}1$–C1–O1
	hNACO(2)	N2–$C^{\alpha}2$–C2–O2
	hNACO(3)	N3–$C^{\alpha}3$–C3–O3
23	hOCNA(1)	O1–C1–N2–$C^{\alpha}2$
	hOCNA(2)	O2–C2–N3–$C^{\alpha}3$

corresponding statistical parameters. These values are used for the prediction of the inter-residue atomic distances and angles by PROGEN.

Structural correlations with the OGP database

Appendix Table 4 shows a typical output (alanine) of the program 'StatisticsBDA.e' with the correlation coefficients, slopes, intercepts and mean values of some of the OGP for different α-carbon distances as the X-variable for alanine residues. The selection of the X-variable was done by considering the correlation coefficients of all the amino acids and determining the best fit. The correlation coefficients for α-carbon distances and 15 geometric parameters for all 20 amino acids are presented in Appendix Table 5.

The correlation coefficients of φ and ψ angles vs. dAA13 were above 0.5 except for a few amino acids. For aspartic acid (D), both the φ and ψ angles were poorly correlated (R-value < 0.5), whereas the ψ angles of all other amino acids were well correlated (R-value > 0.65). The φ angles of glycine (G) and proline (P) did not show any correlation (R-value = 0.02). The φ angle of tryptophan (W) was also poorly correlated (R-value = 0.2). During the positioning of the amino acids in their proper orientations the program PROGEN examined these R-values for the selection of OGP dihedral angles. By trial and error it was found that a R-value cutoff value of 0.3 worked best. If the R-values were above 0.3 the ideal φ and ψ angles were calculated from the values of slope, intercept and mean, but if it they were below 0.3 the contribution of the particular dihedral angle in the trial orientation was disregarded.

An intra-residue dihedral angle, hNACO, (No. 22 in Appendix Table 3) showed a good correlation for most amino acids (> 0.5) with dAA13. This parameter was very helpful in placing the oxygen atoms of the peptide main-chain in their proper position. The intra-residue dihedral angle did not change during the rotation of the amino-acid residue by PROGEN, but the meas-

APPENDIX TABLE 4
OUTPUT FILE OF THE STATISTICAL ANALYSIS PROGRAM FOR THE AMINO ACID ALANINE

AA	X_{var}	Y_{var}	CC	Slope	Intercept	Mean	No.
Ala	dAA13	hPHI	0.66	–42.02	250.00	–83.00	559
Ala	dAA13	hPSI	0.83	112.60	–671.20	28.02	559
Ala	dAA12	dNA	0.34	0.18	0.77	1.47	559
Ala	dAA23	dAC	0.18	0.07	1.24	1.53	559
Ala	dAA23	dCN	0.18	0.10	0.95	1.32	559
Ala	dAA13	aNAC	0.13	0.68	106.94	110.97	559
Ala	dAA23	aACO	0.37	–32.66	244.07	119.92	559
Ala	dAA23	aACN	0.64	57.13	–101.31	115.89	559
Ala	dAA23	aCNA	0.15	15.68	62.69	122.31	559
Ala	dAA23	aOCN	0.34	–24.94	218.85	124.06	559
Ala	dAA23	aFSN	0.11	8.08	–16.16	14.58	559
Ala	dAA23	aFSC	0.49	–19.65	87.59	12.88	559
Ala	dAA23	aSFN	0.06	2.76	–1.71	8.78	559
Ala	dAA23	aSFC	0.44	–31.17	139.45	20.93	559
Ala	dAA13	hNACO	0.66	–100.52	656.68	58.62	559

The optimal AA is Ala 3HHB A:120 with square DVN <0.0001.

urements of some important inter-residue angular parameters (e.g., aOCN and hOCNA) were affected by its correctness. Therefore, this intra-residue dihedral angle was adjusted by rotating C–O along the C^{α}–C axis during the residue positioning. The extent of this rotation was guided by another inter-residue dihedral angle hOCNA, which was highly restricted within a few degrees of its near zero mean value, due to the partial double bond character of the C1–N2 (also C2–N3) linkage of the polypeptide.

It was found that a set of inter-residue angles aFSN, aFSC, aSFN, and aSFC (Nos. 16–19 in Appendix Table 3) of a polypeptide always had small values, within 25°, and showed a good correlation with the second α-carbon distance, dAA23, for a number of amino acids (Table 5). These angles are not bond angles, but they are very helpful in orienting the residues because of their consistently small values. Many other bond angles and distances showed some correlation with one of the three α-carbon atom distances (Appendix Table 5). These correlation coefficients, in conjunction with the slope, intercept and the mean values of regression were used by PROGEN during the positioning of the amino acids in their proper orientations. It was found that a cutoff value of 0.1 gave the best results, instead of 0.3 as in the case of φ and ψ angles. For these parameters, the mean values were taken as the OGP values when the correlation coefficients fell below the cutoff limit and their contributions to the deviation were taken into consideration.

Other parameters such as bond distance dNA and angles aACN showed fairly good correlation

APPENDIX TABLE 5
CORRELATION COEFFICIENTS OBTAINED FROM LINEAR REGRESSION TO IDENTIFY THE DEPENDENCE OF VARIOUS STRUCTURAL PARAMETERS ON ONE OF α-CARBON DISTANCES (dAA12, dAA23, dAA13) FOR EACH AMINO ACID

Y-variable X-variable (dAA)	hPHI 13	hPSI 13	hNA-CO 13	aFSC 23	aFSN 23	aSFC 23	aSFN 23	aNAC 13	aOCN 23	aACN 12	aACO 23	aCNA 23	dCN 23	dNA 12	dAC 23
Ala	0.66	0.83	0.66	0.49	0.11	0.44	0.06	0.13	0.34	0.64	0.37	0.15	0.18	0.34	0.18
Arg	0.66	0.66	0.51	0.45	0.07	0.39	0.02	0.02	0.22	0.61	0.43	0.21	0.28	0.49	0.33
Asn	0.50	0.77	0.19	0.82	0.71	0.84	0.68	0.12	0.10	0.26	0.16	0.02	0.11	0.44	0.11
Asp	0.34	0.47	0.31	0.43	0.10	0.39	0.08	0.17	0.13	0.58	0.50	0.10	0.19	0.51	0.31
Cys	0.51	0.73	0.53	0.47	0.06	0.41	0.01	0.13	0.31	0.61	0.38	0.28	0.27	0.60	0.38
Gln	0.55	0.80	0.69	0.79	0.77	0.80	0.73	0.04	0.02	0.10	0.08	0.02	0.19	0.53	0.18
Glu	0.53	0.74	0.58	0.61	0.47	0.60	0.44	0.19	0.10	0.32	0.29	0.12	0.23	0.49	0.17
Gly	0.02	0.77	0.13	0.38	0.00	0.31	0.04	0.15	0.14	0.53	0.40	0.23	0.36	0.26	0.23
His	0.54	0.68	0.37	0.85	0.81	0.86	0.77	0.09	0.08	0.02	0.09	0.07	0.08	0.51	0.20
Ile	0.65	0.85	0.63	0.34	0.01	0.28	0.04	0.07	0.13	0.50	0.38	0.24	0.28	0.51	0.44
Leu	0.63	0.84	0.58	0.72	0.59	0.73	0.57	0.14	0.11	0.26	0.17	0.01	0.15	0.44	0.20
Lys	0.56	0.71	0.52	0.42	0.04	0.33	0.02	0.03	0.16	0.57	0.44	0.22	0.25	0.44	0.20
Met	0.74	0.92	0.67	0.49	0.17	0.46	0.12	0.06	0.38	0.63	0.24	0.12	0.12	0.31	0.53
Phe	0.58	0.66	0.47	0.87	0.81	0.89	0.79	0.04	0.03	0.19	0.14	0.06	0.09	0.41	0.06
Pro	0.02	0.70	0.63	0.69	0.50	0.69	0.46	0.10	0.14	0.31	0.20	0.03	0.16	0.04	0.16
Ser	0.59	0.72	0.41	0.69	0.46	0.68	0.44	0.10	0.15	0.35	0.22	0.02	0.20	0.47	0.13
Thr	0.61	0.87	0.13	0.38	0.04	0.30	0.06	0.01	0.26	0.57	0.35	0.30	0.17	0.29	0.19
Trp	0.20	0.82	0.57	0.85	0.86	0.87	0.84	0.32	0.03	0.15	0.15	0.06	0.02	0.16	0.00
Tyr	0.54	0.84	0.58	0.90	0.91	0.90	0.89	0.19	0.00	0.03	0.03	0.08	0.13	0.48	0.04
Val	0.61	0.83	0.62	0.44	0.05	0.38	0.02	0.15	0.20	0.58	0.43	0.20	0.13	0.43	0.24

with dAA12 for most of the amino acids (Appendix Table 5). The other distance and angle parameters such as dCN vs. dAA23, dAC vs. dAA23, aNAC vs. dAA13, aOCN vs. dAA23, aACO vs. dAA23, and aCNA vs. aAA23, showed insignificant correlations for most of the amino acids. In calculating the OGP values, the mean values were taken for those with correlation coefficients less than 0.1 and the slopes, intercepts and mean values were used to calculate those with more than 0.1.

The deviations of the bond angles and bond distances from their corresponding OGP values were taken as the differential factors by calculating the ratio of the measured value divided by the ideal value minus 1.0. This was necessary in order to avoid unequal weighting of the parameters of higher values (e.g., the values of bond distances are between 1.0 and 2.0 Å while those of bond angles are around 120°). In the case of the dihedral angles, the ratios of the measured over a fixed value of 90° were used instead of their ideal values; this avoided unequal weighting due to the positions of the mean value in a circular scale which are distributed over a wide range.

Generation of the main chain and refinement of final structure

PROGEN begins orientation of the main chain at the amino-terminus and places the first three amino acids in three positions as shown in Fig. 1. It is assumed that the intra-atomic parameters (see Appendix Table 3, Nos. 6–10, 14, 15 and 22) of the three residues are correct (these values were derived from the statistically OGP of each amino acid). The program then assumes that the orientation of the second amino acid is near perfect and orients the first amino acid in all possible orientations in space with an angular grid of 1.44° around the X-, Y- and Z-axes spanning from −180 to +180°. All the inter-residue parameters between the first and the second residue are calculated and compared in order to determine the r.m.s.d. from the predicted OGP values and the orientation with the least deviation is identified. The orientation of the first residue is close to the correct orientation as compared to the starting arbitrary orientation, but is not necessarily the correct one, because it relies on the orientation of the second residue, which is not at its correct orientation. The program then assumes that this position of the first residue is correct and places the second residue in an orientation where the r.m.s.d. was minimum in the same manner as the first residue. This process is repeated until there is no further significant drop in the r.m.s.d. (usually in 4 cycles). These orientation steps are unique for the very first set of three residues at the amino-terminus of the polypeptide chain. After this step, the second residue is oriented to the minimum r.m.s.d. for the inter-residue parameters in both directions (i.e., first-second and second-third residues) and while keeping the first and the third residues fixed. The third residue is rotated, while the other two remain fixed, and a search for the minimum r.m.s.d. of only one set of parameters between the second and the third residues is made. This process of orienting the second and third residues is repeated until the r.m.s.d. levels off to a small limiting value (about three cycles). After this step, the residues in the three residue positions are moved down one slot, leaving behind the first residue and positioning the next residue in the third slot. The above steps of orienting the second and third residues are then repeated as above, by sliding down the first residue and taking in the next, until the end of the polypeptide is reached.

The number of angular space zones with sides of 1.44° within the whole span of −180 to +180 around X-, Y- and Z-axes is 15.6 million and the calculation of all the inter-residue parameters for these orientations would take an enormously long time. In order to reduce the computational time, the orientations having minimum deviations are searched in three steps, with grid dimen-

sions of (180/5) 36°, (36/5) 7.2° and (7.2/5) 1.44°. Each time the deviation minima is identified, the orientation is re-examined within that space using the next set of smaller grid dimensions. This segmental search reduced the number of orientations for parameter calculations from 15.8 million to 3993 per residue at each orientation step. In addition, the full sphere is not always completely searched during this step. In some instances, only the C–N (No. 2 in Appendix Table 3) distances are calculated and used as the cutoff distance for the calculation of other parameters.

The predicted OGP value of a parameter (Y-variable) is calculated from the value of one of the X-variables (dAA13, dAA12 or dAA23) in conjunction with the slope, intercept, mean and correlation coefficient; this value is stored in the OGP amino acid files (laa*.XYZ) using a subprogram. If the correlation coefficient exceeds a specified limiting value (0.1–0.3), the OGP value is calculated from the slope and intercept. If it is below the specified limit, the mean value is taken as the OGP. During the course of orientation, if the r.m.s.d. does not drop below a limiting value (3.0 Å), the whole process of orientation is repeated with less stringent restrictions. The process of generating the initial set of coordinates is stored in 'Initial_Construct.pdb' and requires about 16 s (total run time) per residue on average using an IRIS 210GTX (Appendix Table 1). These procedures complete 'Step 1' of the construction process.

The input lines 4 through 9 (Appendix Table 1) concern the generation of two QUANTA record files required for the refinement of the initial construct using QUANTA. The FORTRAN programs 'Make_Bond_Break_File' and 'Make_Quanta_Spin_File' generate the record files named 'Break_Unwanted_Bonds.rec' and 'Spin_SideChains.rec', respectively, using 'Initial_Construct.pdb' (the output file on the line 3) as input. The initial ASCII file named 'Initial_Construct.pdb' contains all the atomic coordinates of the polypeptide and is used in QUANTA to refine the configurational and conformational parameters. The pdb file is read by QUANTA to graphically display the molecule (line No. 10 in Appendix Table 1).

The initial refinement of the model (Step 2) involves the rotation of the side chains in order to obtain the best distant-dependent conformation. The side chains are rotated with a cutoff distance of 3.0 Å and an incremental angle of 30°. Rotations are allowed along two single bonds between the α–β and β–γ atoms of all the side chains wherever possible. To automate the process, record files (*.rec) for input to QUANTA are generated a priori using a FORTRAN program named 'Make_Quanta_Spin_File' which takes the pdb file 'Initial_Construct.pdb' as input, and output the file 'Spin_SideChains.rec' containing the QUANTA commands for spinning the side chains (lines 7–9 and 12 in Appendix Table 1). In QUANTA the sulfur atoms form bonds with some neighbouring atoms and they remain intact even when the 'Intra Residue and Named Link' command (under BONDS) is executed. These bonds hinder the rotation of the side chains and therefore, a record file is replayed prior to spinning the side chains in order to break any irrelevant bond between sulfur and other atoms. This record file is generated by the program named 'Make_Bond_Break_File' using the initial pdb file as input and the output is stored in the file 'Break_Unwanted_Bonds.rec' for later use as a QUANTA record file (lines 7–9 and 11).

Further refinement (Step 3) is achieved by the EM of the polypeptide using CHARMm attached to QUANTA (line No. 16 in Appendix Table 1). Before minimization the polar hydrogen atoms are added to the constructed polypeptide using CHARMm (No. 13) and the α-carbon atoms are fixed by applying atom constraints (No. 15). Before applying atom constraints, the name of the molecular structure file (msf) is changed to 'Final_Construct.msf' (No. 14). EM is carried out in three sequential stages, i.e., 100 steps of Steepest Descent, 100 steps of Conjugate

Gradient and 100 steps of Adopted Basis of Newton-Raphson. All the energy terms (i.e., bond length, bond angle, dihedral angle, improper torsion, electrostatic and van der Waals (VDW) energies) are considered during minimization. Nonbond update frequency is set at 50 with a cutoff distance of 8.0 Å. VDW and electrostatic switches are selected as smoothing functions with constant dielectric of value 1.0. Hydrogen bond parameter update frequency is 50 with a cutoff distance of 5.0 Å and a cutoff angle of 90°. Image update frequency is 50 with a cutoff distance of 5.0 Å.

The final refinement (Step 4) is completed using dynamics simulations (No. 17 in Appendix Table 1). The polypeptide is initially 'heated' from 0 K to 400 K, subsequent equilibration at 400 K, followed by a 'quenched' dynamics run from 400 K to 0 K. Temperature and time steps are set at 10 K and 0.0005 ps for heating and 10 K and 0.001 ps steps for quenching. Following the dynamic steps, the EM is repeated (No. 19) using the same CHARMm settings (No. 16). Before minimization, the file is stored as 'Intermediate_Construct' (No. 18). Finally, an ASCII file (in pdb format) containing the refined coordinates of all the atoms of the polypeptide is generated using the 'Export' command in QUANTA (No. 20) and stored in an ASCII file named 'Final_Construct.pdb'.