Reconstruction of protein conformations from estimated positions of the C{alpha} coordinates

P. W. PAYNE

Protein Sci. 1993 2: 315-324

Protein Science (1993), *2*, 315–324. Cambridge University Press. Printed in the USA.

Reconstruction of protein conformations from estimated positions of the C_α coordinates

PHILIP W. PAYNE
Protein Design Labs, Mountain View, California 94043

(RECEIVED August 13, 1992; REVISED MANUSCRIPT RECEIVED October 5, 1992)

Abstract

Protein C_α coordinates are used to accurately reconstruct complete protein backbones and side-chain directions. This work employs potentials of mean force to align semirigid peptide groups around the axes that connect successive C_α atoms. The algorithm works well for all residue types and secondary structure classes and is stable for imprecise C_α coordinates. Tests on known protein structures show that root mean square errors in predicted main-chain and C_β coordinates are usually less than 0.3 Å. These results are significantly more accurate than can be obtained from competing approaches, such as modeling of backbone conformations from structurally homologous fragments.

Keywords: alpha carbon coordinates; computer applications; conformational analysis; knowledge-based methods; potential of mean force; protein secondary structure; protein structure; sequence homology

Accurate principles for building a protein backbone from the C_α coordinates would benefit experimental and theoretical investigations of protein structure. Interpretation of crystallographic electron density maps usually begins with assignment of likely main-chain coordinates, and crystal structures for several important proteins have been published *only* at the C_α coordinate level. Furthermore, knowledge of protein sequences is growing much faster than our ability to purify these proteins and determine their structure by crystallography or NMR. High local homology between the target protein and reference structures occasionally permits transfer of main-chain coordinates and portions of side chains from crystal structures to the target. However, lower levels of homology—corresponding to 30–60% sequence identity—are more common. Side-chain packing variations then preclude direct transfer of side-chain coordinates, but one can still approximate the secondary and tertiary structure of the target protein by transcribing C_α coordinates from the reference structures. This paper addresses the next step—conversion of the C_α coordinates into a complete protein backbone.

This important problem has been widely studied. Constrained energy minimization calculations can accurately rebuild protein backbones from C_α chains, but such computations are time-consuming and sometimes are trapped in incorrect conformations (Levitt, 1983a,b; Brunger, 1988; Kuriyan et al., 1989; Correa, 1990). Another approach searches protein crystal structures for fragments in which the C_α coordinates resemble those of the model being built and then transfers peptide group and side-chain coordinates from the crystallographic reference (Jones & Thirup, 1986; Claessens et al., 1989; Reid & Thornton, 1989; Holm & Sander, 1991; Levitt, 1992). This technique works well in many circumstances but is less reliable for surface loops. Problems in homology modeling of surface loops derive from two sources. First, surface loops are often disordered in the protein crystal structures, so it is hard to obtain accurate coordinates for surface loop templates in the reference crystal structures. Second, surface loops often display greater sequence variability than their buried counterparts. Their conformations can change dramatically to accommodate local differences in hydrophilicity.

This paper takes a fresh approach. The accuracy of homology modeling proves that C_α coordinates contain enough information to locally specify the backbone conformation. But how is this information encoded? Except for occasional cis–trans isomerism, the internal geometry of peptide units is nearly constant: variance of the distance between adjacent C_α atoms is 0.03 Å, and the variance of the amide torsion angle is 4.7°. If C_α positions were known and peptide groups were rigid, the

Reprint requests to: Philip W. Payne, Protein Design Labs, 2375 Garcia Avenue, Mountain View, California 94043.

backbone conformation would be fully specified by rotation of peptide groups about axes that connect adjacent C_α atoms. Neither assumption is strictly true. The C_α coordinates have some statistical uncertainty, and internal coordinates of the peptide groups have small fluctuations. Nonetheless, accurate estimates for backbone coordinates can be obtained by aligning semirigid peptide groups on the C_α framework.

Results

Protein backbones were rebuilt from C_α coordinates by rotating peptide groups about axes that link adjacent C_α atoms. This approach is termed the peptide rotation method. Peptide group positions were chosen to optimize a semiempirical Hamiltonian function, which describes the interaction of adjacent peptide groups, bond angle deformation at C_α, and internal torsion of each main-chain amide bond.

To establish some nomenclature, consider the extended peptide unit: $C_\alpha(k)CONHC_\alpha(k+1)$. The orientation of the peptide group is specified by the dihedral angle between its mean plane and the local plane of the C_α coordinates. Because the extended peptide unit includes only two C_α atoms, there are two ways to define the C_α plane: the third C_α atom precedes or follows the extended peptide group. The first instance defines a peptide orientation angle $W_M(k)$, and the latter defines an orientation angle $W_P(k+1)$. The subscripts M and P, respectively, are mnemonics for minus and plus.

The angle $W_M(k)$ is zero when the carbonyl carbon of residue k is coplanar with and inside the angle formed by C_α coordinates of residues $k-1$, k, and $k+1$. Similarly, the angle $W_P(k+1)$ is zero when the amide nitrogen of residue $k+1$ is coplanar with and inside the angle formed by C_α coordinates of residues k, $k+1$, and $k+2$. The sign of each angle is positive with respect to right-handed rotation of the peptide group about the axis from $C_\alpha(k)$ to $C_\alpha(k+1)$. Both variables measure the orientation of the same peptide group and are related to each other via Equation 1, in which $\tau(k, k+1)$ is the torsion angle of the C_α chain.

$$W_p(k+1) = W_M(k) + \pi - \tau(k, k+1). \quad (1)$$

A Hamiltonian function for protein main chains

The Hamiltonian function of residue k is defined by Equation 2.

$$\begin{aligned} h(k) = {} & E[W_M(k), W_P(k)] \\ & + U[W_M(k), W_P(k), \Omega(k)] \\ & + A\{1 - \cos[2\pi + 2\delta(k)]\}. \end{aligned} \quad (2)$$

The Hamiltonian function for an N-residue protein is the sum of terms for individual residues, except for the initial or final residues, which lack neighboring peptide groups. The leading term of Equation 2 is the potential of mean force (PMF) for nonbonded interactions of two adjacent peptide planes, the second term is the bond angle bending energy, and the final term accounts for torsions about peptide group amide bonds.

Although the Hamiltonian function formally depends on the angles $\{W_P\}$, these are functions of the remaining parameters via Equation 1. The backbone conformation of a protein at residues k and $k+1$ therefore is defined by four variables: the peptide plane orientation angle $W_M(k)$, the amide torsion angle $\delta(k)$, the virtual bond angles of the C_α chain (here denoted as $\Omega(k)$ and $\Omega(k+1)$), and the torsion angle $\tau(k, k+1)$. If peptide groups were rigid, the peptide plane torsion angles and the C_α chain angle would fix the N–C_α–C bond angle. The angle deformation energy likewise would be a function of these variables (Nishikawa et al., 1974; Wako & Scheraga, 1982). Peptide groups in proteins are not completely rigid, but torsional libration about the amide bond nonetheless is small enough that N–C_α–C bond angles computed from the rigid peptide model are reliable.

The standard deviation of peptide torsion angles in our 61-protein database is 4.7°. Internal torsions of the peptide groups therefore have little effect on the PMFs. If one assumes that peptide groups remain rigid the Hamiltonian function of Equation 2 engenders a Boltzmann population distribution $\rho(W_M, W_P, \Omega) = Z^{-1} \exp[-h(W_M, W_P, \Omega)/kT]$, in which Z is the partition function. Integration over possible C_α chain angles yields a distribution function:

$$\begin{aligned} D(W_M, W_P) &\equiv \int d\Omega \rho(W_M, W_P, \Omega) \\ &= Z^{-1} e^{-E(W_M, W_P)/kT} \int d\Omega e^{-U(W_M, W_P, \Omega)/kT}. \end{aligned} \quad (3)$$

The PMF $E(W_M, W_P)$ therefore can be estimated from the frequency distribution $D(W_M, W_P)$ observed in a representative ensemble of protein backbones.

$$\begin{aligned} E(W_M, W_P) = {} & -kT \ln D(W_M, W_P) - kT \ln Z \\ & + kT \ln \int d\Omega e^{-U(W_M, W_P, \Omega)/kT}. \end{aligned} \quad (4)$$

The term proportional to $\ln(Z)$ shifts the energy of each peptide unit by the same constant and will henceforth be ignored. The third term in Equation 4 was evaluated by integrating a harmonic potential function for deformation of the N–C_α–C bond angle. For rigid peptide groups this angle is a function of independent variables W_P and W_M and a single dependent variable, the C_α chain angle Ω.

PMFs were derived by analysis of mean peptide plane orientations in protein crystal structures deposited with the Brookhaven Protein Data Bank (Bernstein et al., 1977). Database statistics were used to evaluate terms on the right-hand side of Equation 4. Details of this derivation are given in the Materials and methods (vide infra). Separate PMFs were derived for glycine, proline, and the other 18 standard amino acids.

Quantitative application

The peptide rotation method was tested on a database of 61 well-resolved, highly refined protein structures. C_α coordinates were extracted from each protein, and peptide group coordinates were chosen to minimize the Hamiltonian function defined by Equation 2. Procedures for optimizing the Hamiltonian and generating coordinates are briefly described in the Materials and methods.

Figures 1 and 2 illustrate the distribution of root mean square (RMS) errors in C_β and carbonyl oxygen positions for the proteins in our test set. A majority of the proteins in our sample have main-chain RMS errors below 0.30 Å, which is comparable to the precision of the crystallographic data. Only 1% of the peptide groups deviates by more than 90°. The high accuracy of predicted side-chain directions will facilitate subsequent generation of complete side-chain models from rotamer libraries (McGregor et al., 1987; Ponder & Richards, 1987; Summers et al., 1987; Islam & Sternberg, 1989) and simulated annealing (Holm & Sander, 1991; Lee & Subbiah, 1991).

Secondary structure

Table 1 shows that the accuracy of coordinates predicted by the peptide rotation method remains high even in the absence of regular secondary structure. The RMS errors in peptide plane angles or C_β positions within turns, omega loops, and irregular secondary structure are comparable to the RMS errors seen in β-ladders. In each case the RMS error of C_β coordinates is less than 0.3 Å, the RMS error of carbonyl oxygen atoms is approximately 0.6 Å, and the RMS error of mean peptide planes is under 30°.

Coordinate errors in α-helices are somewhat smaller than the errors seen in other secondary structure classes. This is somewhat surprising, because the Hamiltonian function used to predict peptide positions makes no provision for hydrogen bonding and was derived solely from irregular secondary structure. However, the regularity expected of a hydrogen-bonded lattice has been indirectly established by provision of C_α coordinates from the crystal structure.

Residue type

Table 2 shows that the accuracy of the rebuilt protein backbones is good for all types of amino acids. Proline, cysteine, and glycine are often treated as special cases in protein conformational analysis, and we have employed distinct potential energy functions for peptide groups ad-

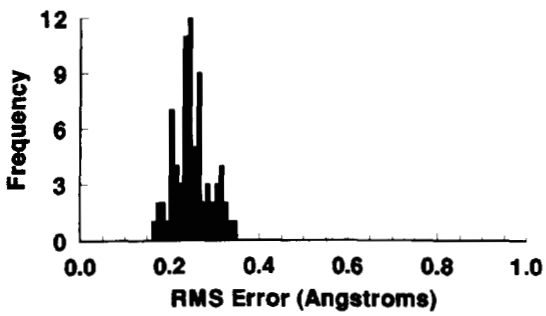

Fig. 1. Root mean square (RMS) errors in predicted side-chain positions. The RMS difference between predicted and crystallographic C_β coordinates was computed for each protein in the test database. Ninety percent of these proteins have C_β RMS errors less than 0.3 Å. Most C_α-C_β bond directions thus are uncertain by less than 15°.

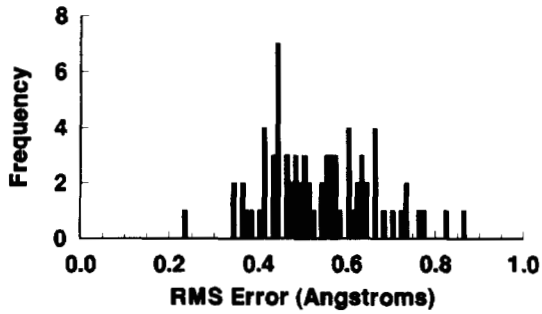

Fig. 2. Root mean square errors in main-chain oxygen coordinates. The RMS difference between predicted and crystallographic coordinates of main-chain oxygen atoms was computed for each protein in the test database. Most of these proteins yield oxygen RMS errors less than 0.6 Å. Errors this small allow recognition of hydrogen-bonded networks that can be refined by subsequent atom-based force fields.

Table 1. *Predicted coordinates are accurate for all secondary structure classes*[a]

Secondary structure	Root mean square error		
	C_β (Å)	O (Å)	W_M (degrees)
α-Helix	0.218	0.423	15.9
Antiparallel β-strand	0.248	0.600	22.5
Parallel β-strand	0.266	0.646	24.8
Reverse turn	0.277	0.633	23.9
Omega loop	0.284	0.659	25.8
Irregular	0.271	0.590	23.1

[a] The root mean square difference between the predicted coordinate and its crystallographic value was calculated using all occurrences of the coordinate in the 63-protein database cited in the text. W_M is the mean peptide plane angle, and O refers to the main-chain oxygen atom.

Table 2. *Main-chain coordinates are accurately predicted for all residue types*[a]

Residue	Root mean square error			Number of occurrences
	C_β (Å)	O (Å)	W_M (degrees)	
Ala	0.211	0.449	16.8	936
Arg	0.260	0.535	19.7	405
Asn	0.276	0.606	22.6	552
Asp	0.295	0.630	25.1	581
Cys	0.259	0.572	21.4	349
Gln	0.240	0.548	19.9	411
Glu	0.265	0.593	24.4	516
Gly	–	0.592	21.8	1,109
His	0.275	0.633	24.2	238
Ile	0.242	0.560	20.8	556
Leu	0.246	0.572	21.3	790
Lys	0.253	0.503	19.3	617
Met	0.277	0.505	18.0	190
Phe	0.279	0.568	22.3	369
Pro	0.236	0.528	21.3	474
Ser	0.257	0.481	18.0	920
Thr	0.268	0.561	21.2	742
Trp	0.262	0.573	21.4	160
Tyr	0.267	0.586	22.1	459
Val	0.240	0.571	22.4	836

[a] Separate potentials of mean force were used for proline and glycine residues. The errors in proline C_β coordinates appear not to be influenced by packing interactions between prolyl and neighboring side chains. The conformational flexibility of glycine does not result in larger coordinate errors.

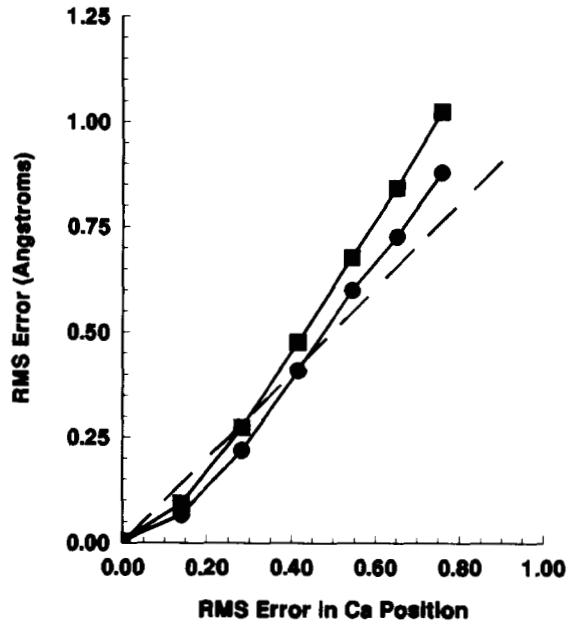

Fig. 3. Stability of backbone generation with respect to errors in the C_α coordinates. Errors in C_β (filled squares) or carbonyl O (filled circles) coordinates are approximately equal to the RMS C_α error below a threshold at 0.8 Å. The C_α positions were randomly displaced subject to a constraint on distances between adjacent C_α atoms. Data sets were generated for various values of the maximum C_α coordinate displacement: 0.25A, 0.50, 0.75, 1.00, 1.25, and 1.50 Å. Each dataset has particular RMS errors for the smeared C_α coordinates and the predicted C_β or O positions. This graph shows the excess error in derived coordinates (C_β or O) attributable to uncertainty in the C_α positions. The excess error is the difference between the observed RMS error and the RMS error incurred when the C_α atoms are at their crystallographic positions.

jacent to a proline or glycine residue. The RMS errors for proline and cysteine are very close to those observed for the entire data set. Furthermore, the conformational flexibility of glycine does not promote larger coordinate errors. The RMS error (0.592 Å) of carbonyl oxygen positions in glycine is close to the mean error (0.556 Å) observed in other residues.

Errors in C_α coordinates

Further application of work reported here is likely to occur in two contexts, crystallographic refinement or de novo modeling of protein substructures. In either case, the C_α coordinates have some intrinsic uncertainty, so it would be useful to know how such errors propagate in the side chain or peptide orientation coordinates. Figure 3 illustrates the excess error in either C_β or carbonyl oxygen coordinates induced by smearing the C_α coordinates. The distances from C_α to either C_β or O of the same residue are nearly constant, so these atoms should approximately follow any displacement applied to the C_α coordinates. This criterion corresponds to the dashed line in Figure 3. The excess RMS error for either C_β or carbonyl oxygen atoms tracks this line well for C_α RMS errors below 0.6 Å. Even above this threshold error amplification is slight. Thus, our backbone reconstruction algorithm is stable for C_α positions uncertain by as much as 1.0 Å.

Discussion

The PMFs

Figure 4 depicts the PMF for general amino acids, which are all residues except proline or glycine. The minimum energy of the β-strand domain at ($W_M = -70°$, $W_P = -130°$) is about 0.7 kcal/mol above the lowest energy of the right α-helical domain at ($W_M = 135°$, $W_P = -110°$). The α-helical and β-strand domains are separated by a relatively low barrier, which is 1.5 kcal/mol above the floor of the α-helical well and only 0.8 kcal/mol above the floor of the β-strand energy well. A third energy well centered near ($W_M = -135$, $W_P = 110$) corresponds to left-helical conformation.

This PMF provides some insight into dynamics of peptide motion during protein folding. When the peptide chain ends at residue $k + 1$ there is a populated range of C_α chain angles $\Omega(k)$ that allows the peptide plane be-

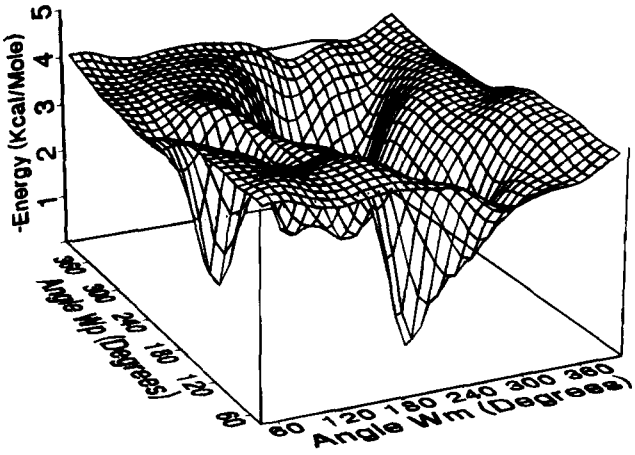

Fig. 4. Potential of mean force for peptide planes that meet at a residue other than glycine or proline. W_M is the dihedral angle formed by the mean peptide plane in the C-terminal direction. W_P is the dihedral angle of the mean peptide plane in the N-terminal direction. The global energy minimum near ($W_M = 150°$, $W_P = 250°$) has (right) α-helical conformation. A second energy well, which has β-sheet conformation, is centered near ($W_M = 290°$, $W_P = 230°$). The floor of this energy well is about 0.7 kcal/mol above that of the α-helical domain. The energy minimum near ($W_M = 210°$, $W_P = 110°$) belongs to a left α-helix.

tween residues k and $k + 1$ to flip between α-helical and β-strand conformations. As the peptide plane flips, the orientation angle W_M of the peptide group connecting residues k and $k + 1$ changes by approximately 160°, but the orientation angle W_P of the neighboring peptide group changes little. In each of the three principal energy wells the angle $W_P(k)$ is constrained to a range no more than 60° wide. Extension of the peptide chain from residue $k + 1$ to $k + 2$ therefore locks the previously mobile peptide group connecting residues k and $k + 1$ into one of the three local conformations.

Although protein transcription proceeds from the N-terminus to the C-terminus, the foregoing discussion does not claim that protein folding occurs strictly in order of transcription. There is ample evidence for multiple kinetically controlled folding pathways, some of which require nucleation centers midway through the protein sequence (Creighton, 1992). But our analysis of the PMFs does show that packing interactions that immobilize the side chain of residue $(k + 1)$ or the following peptide plane also rigidify the secondary structure of the immediately preceding residues. Such processes may help explain transient stabilization of folding intermediates.

The minimum energy of the left α-helix is about 0.8 kcal/mol above that of the right helix, and the walls of the energy well are much steeper. All pathways leading to the left-helical conformation of general residues traverse a barrier greater than 3.5 kcal/mol. This conformation is expected only in strained environments or when a sharp twist of the backbone is needed for a particular tertiary fold. In the latter instance, the steep walls surrounding this minimum provide especially sharp definition of the peptide plane and side-chain orientations.

The PMF physically represents the electrostatic and van der Waals interactions of adjacent peptide groups and steric contacts of these peptide groups with the side chain of their common residue. Because glycine lacks a C_β atom one can reflect the adjacent peptide groups through the local C_α plane without changing any of the atomic distances that underlie the PMF. The glycine PMF therefore should have inversion symmetry with respect to W_M and W_P. However, the crystallographic population density is highly asymmetric. The population statistics were therefore symmetrized with respect to inversion of W_M and W_P before the PMF was computed.

The glycine PMF has extrema at ($W_M = 150°$, $W_P = 260°$) and ($W_M = 210°$, $W_P = 100°$), which respectively fit right α-helical and left α-helical structures. The well-known glycine turn has peptide plane orientation angles in the left α-helical domain. Another local energy minimum near ($W_M = 330°$, $W_P = 260°$) is a β-strand conformation; its minimum energy lies about 0.3 kcal/mol above the energy floor for the helical conformers. A fourth energy well near ($W_M = 30°$, $W_P = 100°$) is the mirror image of this β-strand conformation. Although it is thermodynamically accessible, it is rarely occupied in protein crystal structures. This discrepancy suggests that peptide plane orientations at glycine sometimes are influenced by thermodynamic coupling to nearby residues. The glycine PMF also features a broad plateau near ($W_M = 0°$, $W_P = 0°$). The energy on this plateau is about 2.0 kcal/mol above the PMF in stable conformations. Hence, protein backbones bend at glycine when such distortion contributes at least 2.0 kcal/mol to stabilizing interactions elsewhere in the protein.

The PMF for *trans* proline has steep local minima near ($W_M = 140°$, $W_P = 250°$) and ($W_M = 305°$, $W_P = 225°$). The first well is an α-helical conformation; the other is a β-strand conformation. Both wells have similar depth, and a barrier exceeding 4 kcal/mol hinders interconversion of the two conformers.

Comparison with other de novo approaches

Other methods for rebuilding protein backbones from the C_α coordinates fall broadly into two classes. Like the present work, the de novo techniques apply predefined formulas to an isolated C_α chain. In contrast, the homology-based models search a database of protein crystal structures for segments that have C_α chain conformations similar to the C_α coordinates of the protein being modeled. The de novo methods are considered first.

Constrained molecular dynamics has been applied to regeneration of backbone coordinates for alpha-lytic protease (Brookhaven File 2ALP) from an initial set of C_α coordinates (Correa, 1990). The main-chain RMS error was about 0.1 Å less than that reported here, but this im-

provement required 14 h on a 50-MHz array processor. In contrast, the peptide rotation method rebuilt the backbone of this protein in less than 30 s using a Silicon Graphics 310 workstation. The faster speed of the peptide rotation method would recommend its use in conformational analysis of protein surface loops because thousands of competing C_α templates must be considered.

Others have rebuilt protein main chains by using residue-dependent mean direction cosines to align each C_α–C_β vector with respect to the local C_α plane (Rey & Skolnick, 1992). Peptide group coordinates were then determined by minimizing a penalty function that describes deviations of bond angles and bond lengths from idealized values. Resulting RMS errors in main-chain coordinates (see Table 3) were substantially worse than those reported here, and carbonyl oxygen errors were twofold greater than those obtained here.

Rotation of peptide groups about the axes between adjacent C_α atoms has been considered by other investigators (Nishikawa et al., 1974; Wako & Scheraga, 1982) who derived equations that relate ϕ and ψ angles of adjacent residues to geometric parameters of the C_α chain. Such relationships have been used to rebuild bovine pancreatic trypsin inhibitor (BPTI) from C_α coordinates (Purisima & Scheraga, 1984). Unfortunately, the assumed rigidity of bond angles and amide torsion angles led to numerical instabilities that precluded prediction of the BPTI structure. The present work avoided similar problems by explicitly incorporating angle relaxation in the Hamiltonian.

Luo et al. (1992) also rebuild protein backbones by rotating peptide groups about the virtual axes that connect successive C_α atoms. They accept all peptide plane orientations that yield N–C_α–C bond angles within 5° of the idealized tetrahedral angle. Sets of main-chain coordinates are scored by calculating the number of ϕ and ψ angles that fall within bounds suggested by statistical analysis (Moult & James, 1986) of the Protein Data Bank. In contrast, the present study has converted database statistics to PMFs. The availability of potential energy functions for peptide plane interactions, bond angle bending, and peptide plane torsion helps our method redistribute local strain energy. In most protein structures a few bond angles or main-chain dihedral angles adopt extreme values in order to minimize the energy of the entire main chain. This is especially likely near sharp bends in the C_α trace. Our method accurately predicts occurrence of these atypical coordinates and yields RMS coordinate errors about half the magnitude of those reported by Luo et al. (1992).

Comparison with homology-based approaches

Based on the pioneering work of Jones and Thirup (1986) and Blundell and coworkers (1988), several investigators have rebuilt protein backbones from C_α coordinates by splicing fragments culled from accurate protein crystal structures. Table 3 compares the quality of protein backbones predicted by these methods, which we will collectively describe as segment match modeling (Levitt, 1992), and backbone models built as described in the present work. All segment match algorithms follow a common theme. A reference database is searched to identify clusters of C_α atoms that have the best RMS overlaps with a group of C_α atoms in the target structure. When a good match is found, some, or perhaps all, of the coordinates in a segment are copied from the database into the target structure.

The precise rules for choosing segments and smoothing discontinuities at segment boundaries vary widely. Flavodoxin was rebuilt by manually selecting the segments spanning 4–7 residues that have best overlap with C_α atoms of homologous segments in the target chain (Reid & Thornton, 1989). The main-chain RMS error is 0.51 Å, but the errors in turns (0.69 Å) and zones of irregular secondary structure (0.77 Å) are substantially worse.

Table 3. *Accuracy of backbone coordinates predicted by the peptide rotation method*[a]

	Main-chain RMS error (Å)		
Protein	This paper	Prior work	Citation
2ALP	0.30	0.19	Correa, 1990
2APP	0.30	0.37	Levitt, 1992
5CPA	0.33	0.61	Claessens et al., 1989
1CRN	0.20	0.56	Levitt, 1992
1CTF	0.19	0.29	Levitt, 1992
2CTS	0.33	0.54	Claessens et al., 1989
4FXN	0.39	–	–
3FXN	–	0.51	Reid and Thornton, 1989
	–	0.49	Correa, 1990
	–	0.77	Rey and Skolnick, 1992
	–	0.44	Levitt, 1992
1GCR	0.26	0.40	Holm and Sander, 1991
2MHR	0.22	0.78	Rey and Skolnick, 1992
2PRK	0.26	0. 49	Holm and Sander, 1991
6PTI	0.32	0.51	Levitt, 1992
4PTI	–	0.68	Rey and Skolnick, 1992
1TIM	0.50	0.55	Claessens et al., 1989
	–	0.68	Rey and Skolnick, 1992
3TLN	0.32	0.38	Levitt, 1992
1UBQ	0.25	0.42	Holm and Sander, 1991
3WGA	0.33/0.29	0.48	Holm and Sander, 1991
2WRP	0.18	0.46	Holm and Sander, 1991

[a] Most of the examples listed under Prior work used fragment homology to build main-chain coordinates. The root mean square (RMS) error in main-chain coordinates includes deviations of C_α, N, C, and O atoms. Proteins are identified by their Brookhaven Protein Data Bank codes. Note that WGA has two chains. Our algorithm generates a separate RMS error for each chain. TIM has substantially poorer resolution (2.5 Å) than the other structures and is included here only for the purpose of comparing our results with prior work. TIM was not included in the database from which the potentials of mean force were built.

The reconstruction of cellobiohydrolase II (Jones et al., 1991) used overlapping 5-residue segments and yielded RMS errors of 0.56 Å for the main chain and 1.0 Å for the carbonyl oxygen atoms. There was a propensity for peptide flips at sites without good matches between the target C_α coordinates and segments from the crystallographic database. This is in accord with our observation that peptide flips tend to occur at sites where main-chain bond angles are deformed; these deformities will be hard to reproduce in the crystallographic database, so one would expect a correlation between peptide flipping and the occurrence of high RMS coordinate differences between the target C_α segment and homologous database entries.

Other investigators prefer to use the longest database fragments that have C_α RMS error less than 0.5 Å (Claessens et al., 1989). To reduce conformational discontinuities at segment boundaries these investigators overlapped the segment search so that each selected segment has three residues in common with its predecessor. This protocol generated a pool of 50 reasonable crystallographic segments in each zone. The best-fitting member of each set was used for backbone regeneration.

The main-chain RMS errors obtained for proteins 2CTS, 5CPA, and 1TIM all exceeded 0.5 Å; this is not surprising because the Claessens segment search was designed to produce pool members quickly, but with accuracy not much better than 0.5 Å. Although overall RMS statistics are reasonable this algorithm performs poorly in loops (1.3 Å RMS), and the secondary structure is unstable at segment boundaries. There is a far more serious practical problem with implementation of this algorithm. As proof of principle, Claessens et al. (1989) rebuilt protein backbones from the best-fitting member of each pool. But this criterion cannot be followed in practice because there is no way to define the best match without knowing the answer in advance.

To overcome this difficulty, other investigators (Holm & Sander, 1991) first generated a series of 50 best-matching segments centered at each residue junction and then employed dynamic programming to select the most compatible pairs of overlapping segments. Their main-chain RMS errors are approximately 0.2 Å greater than those found via our method. Nearly 10% of the peptide groups are flipped in proteins built by the Holm–Sander method.

A particularly sophisticated implementation of homology modeling has recently been described (Levitt, 1992). It begins with enumeration of 40 database segments, each 3–4 residues long, which have good RMS fit to known coordinates in the target structure. Such segment sets are built in the neighborhood of every residue. Each segment has an effective energy that is defined as a weighted average of the RMS distance error and the nonbonded interaction energy between the segment and its environment. Segments are combined by a stochastic procedure that uses Metropolis sampling to select coordinates from four lowest energy segments at each residue. Averaging of coordinates over the resulting ensemble generates an initial guess for the protein backbone. These coordinates are refined by a subsequent energy minimization step.

The six examples in Table 3 show that this procedure is more accurate than the other algorithms that rebuild a protein backbone using fragments from the crystallographic database. In some cases (1CTF, 3TLN, 4FXN, 2APP) the main-chain RMS errors are close to (but larger than) the main-chain RMS coordinate errors obtained with our algorithm. But in two cases (1CRN, 6PTI) our methodology is significantly more accurate.

Conclusions

In summary, the high accuracy of our predictions shows that our model incorporates most of the physical factors that govern protein backbone conformations. In particular, cooperative relaxation of peptide plane rotation angles and backbone bond angles is important because peptide planes and C_α bond angles frequently assume nonideal values to optimize the cumulative chain energy. Computational implementation of this model facilitates very rapid reconstruction of protein backbone conformations. The reliability, computational simplicity, and conceptual foundations of our model should improve crystallographic refinement and modeling of new proteins and their surface loops.

In its present form our algorithm requires estimates of coordinates for the C_α atoms. Errors in the C_α coordinates are replicated in the final structure. Database methods that splice homologous fragments are able to supply the missing C_α coordinates and suggest corrections to possibly anomalous atom positions. For these reasons they remain more robust than the present work. Nonetheless, the PMFs described in this report incorporate most of the physical interactions relevant to protein backbone conformations. Our laboratory is currently using such functions for ab initio determination of C_α coordinates for short polypeptide fragments embedded in proteins. The latter study will be the subject of a future report.

Materials and methods

Database selection

PMFs were derived by analysis of mean peptide plane orientations in a subset of protein crystal structures deposited with the Brookhaven Protein Data Bank (Bernstein et al., 1977). The initial pool comprised entries with resolution below 2.0 Å and *R*-factor below 0.20. Because the immunoglobulins were underrepresented in this pool, structures 2FB4 (KOL) and 3FAB (NEW) were included even though their quality is somewhat lower.

When a particular protein or its mutants was represented by several high-resolution structures only one structure was kept. Thus, the database kept entries 2LZM (T4 lysozyme), 2SGA (proteinase A from *Streptomyces griseus*), and 4PTP (beta trypsin) but discarded the respective matching structures 1L01, 1SGC, and 2PTN. Proteins that were refined using idealized molecular geometry or that were postprocessed by simulated annealing were removed from the dataset. Several entries (3C2C, 4DFR, 4TNC, 5PTI) were rejected because the standard error of either the adjacent C_α distances or the main-chain bond angles at C_α was significantly higher than the norms derived from the database as a whole.

The validated database comprised the following 61 Brookhaven entries: 1BP2, 1CSE, 1CRN, 1ALC, 1CTF, 1GCR, 1GD1, 1GOX, 1GP1, 1HMQ, 1HOE, 1LZ1, 1MBC, 1MLT, 1PAZ, 1PPT, 1RDG, 1SGT, 1TON, 1UBQ, 2ACT, 2ALP, 2APP, 2APR, 2AZA, 2CDV, 2CI2, 2CPP, 2FB4, 2HHB, 2LHB, 2LZM, 2MHR, 2OVO, 2PRK, 2RHE, 2SGA, 2UTG, 2WRP, 3EBX, 3EST, 3FAB, 3GRS, 3INS, 3RNT, 3RP2, 3SGB, 3TLN, 3WGA, 451C, 4FD1, 4FXN, 4PTP, 5CHA, 5CPA, 5CYT, 6PCY, 6PTI, 7RSA, 8DFR, 9PAP. No attempt was made to eliminate homologous proteins. This dataset contains two lysozymes and several serine proteases. That would be a source of error if one were studying statistical patterns in protein tertiary structure. However, this work uses the database to derive local conformational properties that span at most three contiguous residues, so homologous folds or active sites are not a significant problem. The dataset contains 9,476 residues.

Angle deformation and torsional potentials

Bond angle strain

The main-chain Hamiltonian function defined by Equation 2 includes terms for bond angle bending and torsions about peptide group amide bonds. The harmonic deformation potential of the N–C_α–C bond angle is $V(\theta) = c(\theta - \bar{\theta})^2$, where c is 66.1 kcal/rad^2 and $\bar{\theta} =$ 1.934 radians. The force constant was derived from linewidths of Gaussian distributions for main-chain bond angles in our 61-protein dataset and is close to the force constant used by other investigators for peptide and protein molecular mechanics calculations.

The second term in Equation 2 indicates that the bond angle strain energy is an implicit function of the peptide plane torsion angles and the virtual bond angle of the C_α chain. Equation 5 defines the relationship between the N(k)–C_α(k)–C(k) bond angle, denoted here by θ, and the mean peptide plane orientation angles. Denote the virtual bond angle $C_\alpha(k-1)$–$C_\alpha(k)$–$C_\alpha(k+1)$ by Ω, the N(k)–C_α(k)–C_α($k-1$) bond angle by δ, and the C(k)–C_α(k)–C_α($k+1$) bond angle by γ. Straightforward vector analysis yields the desired formula for θ.

$$\begin{aligned}\cos(\theta) = {} & \cos\Omega[\cos\gamma\cos\delta - \sin\gamma\sin\delta\cos W_M\cos W_P] \\ & + \sin\Omega[\sin\gamma\cos\delta\cos W_M \\ & \qquad + \cos\gamma\sin\delta\cos W_P] \\ & + \sin\gamma\sin\delta\sin W_M\sin W_P.\end{aligned} \tag{5}$$

Use of the McLaurin series for the cosine function, expanded about $\bar{\theta}$, yields the following approximation for the bond angle bending energy.

$$V(\theta) \cong 2c[1 - \cos(\theta - \bar{\theta})]. \tag{6}$$

Because cos θ is a function of W_M, W_P, and Ω, the angle strain energy likewise is a function of the latter variables:

$$U(W_M, W_P, \Omega) \equiv V[\theta(W_M, W_P, \Omega)]. \tag{7}$$

Nonplanar peptide groups

The third term in Equation 2 accounts for torsional puckering of peptide groups about the N–C bond. The torsional potential originates from electronic conjugation between the amide lone pair and the carbonyl group π orbitals and therefore is described by a twofold barrier:

$$E(\omega) = A[1 - \cos(2\omega)]. \tag{8}$$

For trans peptide groups one may write $\omega = \pi + \delta$. Taylor expansion of Equation 8 for small δ yields the harmonic formula $E(\delta) = 2A(\delta)^2$, and substitution of this energy into the Boltzmann law predicts a Gaussian distribution for δ. The standard deviation of δ in our 61-protein database (4.68°) thus corresponds to a value of 22.5 kcal rad^{-2} for the force constant A. These statistics discount the few outliers for which $|\delta| > 16.0°$. The analogous expression for a cis peptide group omits the phase shift, 2π. The same rotational barrier height is used for cis and trans peptide groups.

Potentials of mean force

Separate PMFs were derived for glycine, proline, and the other 18 standard amino acids. Members of the latter set are called general residues because they all are alanine substituents and should display similar main-chain conformational properties. All secondary structure classes were bundled in the PMFs for glycine, based on 908 residues, and proline, based on 342 cases. The PMF for general amino acids was built from 2,277 residues without regular secondary structure (Kabsch & Sander, 1983). Nevertheless, control data sets were generated for α-helices or β-sheets. Remarkably, density functions $D(W_M, W_P)$ built solely from regions with irregular secondary structure are statistically equivalent to composites of population distributions generated from α-helices or β-sheets.

Using parameters derived from irregular secondary structure to describe regions possessing regular secondary structure may seem unusual. However, the Ramachandran map, $E(\phi,\psi)$, contains broad energy minima near the torsion angles observed in idealized α-helices or β-sheets. Residues that do not participate in regular secondary structure are nonetheless subject to Ramachandran constraints and often have α-like or β-like backbone conformations. The PMFs for residues in irregular secondary structure therefore ought to resemble the composite energy for regions that have regular secondary structure.

Each data pair (W_M, W_P) from our 61-protein dataset was replaced by a two-dimensional Gaussian function. Exponents of these Gaussian functions were optimized to fit the local density of states. The sum of all such Gaussian functions generated a continuous, positive density distribution, $D(W_M, W_P)$. Conformations were sampled at 5° intervals, and the density function was also tabulated at this resolution.

The PMFs for proline, glycine, and general amino acids were obtained by substituting the appropriate density distributions $D(W_M, W_P)$ into Equation 4. They respectively have two, four, and three dominant energy wells. In addition there are numerous local extrema in the PMFs. The latter local minima are not physically significant because they span a range less than 0.5 radians wide, typically are shallow, and have relative energies at least 2.5 kcal/mol above the global energy minimum. They typically occur in regions where the source data set was sparse.

To eliminate noise associated with spurious local minima, the PMFs were fit with analytic functions that describe line shapes of the principal energy wells. The line shape of each well was expanded using a product of [0, 4] Pade approximations in variables ν_1 and ν_2. These variables are periodic functions of W_M and W_P, and they are chosen to diagonalize the principal axes of each well. A majority of points in the PMF can be fit with precision under 0.3 kcal/mol, and nearly all points that have energy under 4.0 kcal/mol are fit with estimated errors under 0.5 kcal/mol.

Optimization of the main-chain Hamiltonian function

The conformation and Hamiltonian function of a protein backbone are defined by four variables per residue: the peptide plane orientation angle $W_M(k)$, the internal peptide twist angle $\delta(k)$, the virtual bond angle $\Omega(k)$, and the C_α chain torsion angle $\tau(k-1, k)$. Because this paper is investigating reconstruction of complete protein backbones from crystallographic C_α coordinates the Ω and τ values are known. The only adjustable variables in our model are the mean peptide plane angles $W_M(k)$ and twist angles $\delta(k)$. These are chosen to minimize the sum of residue Hamiltonian functions defined by Equation 2.

Because peptide groups interact only through nearest-neighbor potentials, dynamic programming efficiently identifies the global energy minimum of the Hamiltonian function. This work employed discrete tables of W_M and values δ for each residue. Peptide plane orientation angles W_M were initially scanned in 5° steps while keeping peptide groups planar. The best conformation from this search was refined by varying the W_M values in 1.5° steps over a range of 7.5° above or below the proposed solution. During the second search, peptide groups could become nonplanar; torsion angles $\delta(k)$ were adjusted in 2° steps to minimize the Hamiltonian function.

Coordinate generation

Because peptide groups interact only through nearest-neighbor potentials, dynamic programming efficiently generates lists of peptide plane orientation angles $\{W_M\}$ and amide torsion angles $\{\delta\}$ that minimize the Hamiltonian function. Conversion of the peptide plane orientations to atomic coordinates employs standard values for bond angles and bond lengths. Mounting peptide groups on a fixed C_α chain is possible only if other internal coordinates of the peptide groups respond to fluctuations of the amide torsion angles. Thus, internal bond angles of each peptide group were both raised or lowered by a common amount γ (approximately 1°) to maintain a fixed distance between C_α atoms as $\delta(k)$ changed. The relaxed peptide group Cartesian coordinates were aligned with the axis that joins atoms $C_\alpha(k)$ and $C_\alpha(k+1)$ and were rotated as specified by $W_M(k)$. Finally, the C_α–C_β bond vectors were chosen to yield tetrahedral stereochemistry at C_α.

References

Bernstein, F., Koetzle, T., Williams, G., Meyer, E.F., Jr., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). The Protein Data Bank: A computer-based archival file for molecular structure. *J. Mol. Biol. 112*, 535–542.

Blundell, T., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T., Overington, J., Singh, D., Sibanda, B., & Sutcliffe, M. (1988). Knowledge-based protein modelling and design. *Eur. J. Biochem. 172*, 1179–1188.

Brunger, A. (1988). Crystallographic refinement by simulated annealing. Application to a 2.8 Ångstrom resolution structure of aspartate aminotransferase. *J. Mol. Biol. 203*, 803–816.

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

Correa, P. (1990). The building of protein structures from α-carbon coordinates. *Proteins Struct. Funct. Genet. 7*, 366–377.

Creighton, T.E., Ed. (1992). *Protein Folding*. W.H. Freeman & Co., New York.

Holm, L. & Sander, C. (1991). Database algorithm for generating protein backbone and side-chain coordinates from a Cα trace. Application to model building and detection of coordinate errors. *J. Mol. Biol. 218*, 183–194.

Islam, S. & Sternberg, M. (1989). A relational database of protein structures designed for inquiries about conformation. *Protein Eng. 2*, 431–442.

Jones, T.A. & Thirup, S. (1986). Using known substructures in protein model building and crystallography. *EMBO J. 5*, 819–822.

Jones, T., Zhou, J., Cowan, S., & Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A47*, 110–119.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: Pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*, 2577–2637.

Kuriyan, J., Brunger, A., Karplus, M., & Hendrickson, W. (1989). X-ray refinement of protein structures by simulated annealing: A test of the method on myohemerythrin. *Acta Crystallogr. A45*, 396–409.

Lee, C. & Subbiah, S. (1991). Prediction of protein side-chain conformations by packing optimization. *J. Mol. Biol. 217*, 373–388.

Levitt, M. (1983a). Molecular dynamics of native proteins: Computer simulation of trajectories. *J. Mol. Biol. 168*, 595–620.

Levitt, M. (1983b). Protein folding by constrained energy minimization and molecular dynamics. *J. Mol. Biol. 170*, 723–764.

Levitt, M. (1992). Accurate modeling of protein conformations by automatic segment matching. *J. Mol. Biol. 226*, 507–533.

Luo, Y., Jiang, X., Lai, L., Qu, C., Xu, X., & Tang, Y. (1992). Building protein backbones from Cα coordinates. *Protein Eng. 5*, 147–150.

McGregor, M., Islam, S., & Sternberg, M. (1987). Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J. Mol. Biol. 198*, 295–310.

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

Nishikawa, K., Momany, F., & Scheraga, H. (1974). Low-energy structures of two dipeptides and their relationship to bend conformations. *Macromolecules 7*, 797–806.

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

Purisima, E. & Scheraga, H. (1984). Conversion from a virtual-bond chain to a complete polypeptide backbone chain. *Biopolymers 23*, 1207–1224.

Reid, L. & Thornton, J. (1989). Rebuilding flavodoxin from Cα coordinates: A test study. *Proteins Struct. Funct. Genet. 5*, 170–182.

Rey, A. & Skolnick, J. (1992). Efficient algorithm for the reconstruction of a protein backbone from the α-carbon coordinates. *J. Comp. Chem. 13*, 443–456.

Summers, N., Carlson, W., & Karplus, M. (1987). Analysis of side-chain orientation in homologous proteins. *J. Mol. Biol. 196*, 175–198.

Wako, H. & Scheraga, H. (1982). Distance-constraint approach to protein folding. II. Prediction of the three-dimensional structure of bovine pancreatic trypsin inhibitor. *J. Protein Chem. 1*, 85–117.