# Efficient Algorithm for the Reconstruction of a Protein Backbone from the α-Carbon Coordinates

**Antonio Rey* and Jeffrey Skolnick†**

*Department of Molecular Biology, Scripps Research Institute, La Jolla, California 92037*

*Received 13 August 1991; accepted 21 October 1991*

We present an analytical method for generating a whole protein backbone structure from the coordinates of the $\alpha$-carbons. The procedure begins by automatically positioning the $\beta$-carbons for every residue, and then the positions of the carbonyl groups and the amide nitrogens are also computed. The method is based upon the simultaneous minimization of a number of geometrical constraints that appear in real proteins and that can be very easily formulated as a set of trigonometric relations between the coordinates of the atoms involved in the backbone reconstruction. The resulting algorithm has been tested for proteins of very different sizes and topologies, and can advantageously compete with other methods proposed for this goal both in accuracy and in computational requirements. Possible ways of further refinement of the resulting structures are discussed.

## INTRODUCTION

During the past several years, the reconstruction of the full-atom representation of a protein from the coordinates of its $\alpha$-carbons has received a lot of attention. There are a number of reasons for this. First, the map of the $\alpha$-carbons is usually generated in the early stages of the solution of a new crystal structure studied through X-ray diffraction. In addition, a fraction of the structures compiled in the Brookhaven Protein Data Bank[1] only contain the $\alpha$-carbon coordinates. Finally, but most important, protein modeling techniques for the prediction of the structure of proteins are continuously employed.[2] In this context, it is useful to have a series of tools that allow one to pass from a simplified representation, usually based only on an $\alpha$-carbon trace, to more sophisticated models that include all atoms contained in the protein backbone and, if possible, the side chains, thereby allowing for the study of more detailed, specific processes.

Most procedures usually employed in the full-atom reconstruction are based upon different kinds of "backbone dictionaries."[3-5] The sequence of $\alpha$-carbons is split into several fragments, and for every one of them a series of well-refined structures included in a certain crystallographic database is scanned for patterns with the same (or very similar) topology. These matching patterns are used as templates for the placement of additional atoms. One usually encounters a problem when trying to join the different fragments and close the possible remaining gaps. This is a difficult procedure that usually requires several complicated refinement steps.

On the other hand, some methods avoid the use of statistical databases by making a refinement of the structure along with its construction,[6] i.e., an energy minimization procedure is run after positioning every full-atom amino acid (with the side chains restricted to the $\beta$-carbon in the first stage) so the protein is sequentially rebuilt. This method is capable of providing accurate final structures, although it seems to be rather expensive from the computational point of view, especially for proteins containing a medium to large number of residues.

One additional point that needs to be taken into consideration is the problem that arises when $\alpha$-carbon coordinates do not exactly correspond to the set of distances observed in the existing database of crystal structures. A clear example appears in the theoretical studies of dynamical processes that use a lattice projection of the protein.[7] Even by choosing a lattice spacing consistent with the average distance between $\alpha$-carbons found in real proteins the geometry of the backbone is clearly distorted so methods based on the overlapping of fragments of real proteins are essentially useless under these conditions. Nevertheless, it would be useful to have a method that would allow for a full-atom reconstruction of the protein for these models, both as a test of the quality of the lattice representation and to

*Permanent address: Departamento de Química Física, Facultad de Ciencias Químicas, Universidad Complutense de Madrid, E-28040 Madrid, Spain.

†Author to whom all correspondence should be addressed.

CCC 0192-8651/92/040443-14$04.00

provide the possibility for further improvements of the model.

In this article, we propose a method for building the backbone atoms of a protein, including the $\beta$-carbons, from any set of real or reasonable approximate $\alpha$-carbon representation. An important difference between this method and those mentioned previously is that the procedure has a purely analytical basis, resulting from the trigonometric relations existing between the different bond lengths and bond angles involved in the protein backbone. The possible ambiguity arising from the rotation around single bonds (or, in other words, from the different values of the torsional angles $\phi$ and $\psi$ in the backbone) disappears when the positions of the $\alpha$-carbons are known. In addition, the possible existence of multiple solutions to the problem as a consequence of the rotation of the peptide plane around the axis that joins two adjacent $\alpha$-carbons also vanishes once the positions of the $\beta$-carbons have been determined.

This is not the first time an analytical method has been proposed for backbone reconstruction. Almost 10 years ago, one other method was developed for this purpose.[8] This method was formulated in terms of the torsional angles $\phi$ and $\psi$, instead of the Cartesian coordinates we shall employ here, and propagated the reconstruction of backbone atom positions from an initial guessed estimation of a ($\phi_i$, $\psi_i$) pair for an inner residue. While this method provided excellent results for ideal rigid protein geometries, its application to an $\alpha$-carbon trace taken from a real protein did not yield a very good fit, especially due to the divergence between the real $\alpha$-carbon trace and the ideal one. As shown below, our method is not so dependent on the rigid geometry assumption (although, due to the fixed bond lengths and angles, it provides the best results for idealized structures), and does not depend on any initial guess for the varying geometrical parameters along the chain.

## DETERMINATION OF β-CARBON POSITIONS

The first step in our reconstruction method, and in fact a remarkable result on its own, is the possibility of determining the coordinates of the $\beta$-carbon positions from the $\alpha$-carbon trace of the protein. As a matter of fact, the feasibility of this procedure is not a surprise. If statistical methods based upon the scanning of databases can build the whole protein, with the multiplicity of torsional angles involved, there must be geometrical (or chemical) reasons that allow the direction of the bond joining $\alpha$- and $\beta$-carbons for every residue to be fixed. The only problem, then, is to find an adequate reference system that allows the definition of a series of geometrical constraints that uniquely position the $\beta$-carbon. Obviously, any external frame is discarded since the different global orientations of the protein do not have any effect on the relative position of the atoms. The natural reference system for this problem should be centered in the $\alpha$-carbon of the considered residue, with the axes directed according to the local backbone geometry. Thus, to position the $\beta$-carbon of the $i$th residue, we have defined the reference system in the following way (see Fig. 1): The first axis, $\mathbf{u}_1$, is computed in the direction of the cross-product of the two vectors that join atoms $C_i^\alpha$ with $C_{i+1}^\alpha$ and $C_i^\alpha$ with $C_{i-1}^\alpha$, named as $\mathbf{r}_{i,i+1}$ and $\mathbf{r}_{i,i-1}$, respectively, and is thus perpendicular to the plane defined by these three atoms (which can never be found in a linear configuration neither in a real protein nor in any valid model representation). The second axis of the reference system, $\mathbf{u}_2$, is defined by the direction opposite to the sum of the same two vectors normalized to unit length. This second vector, consequently, is in the plane of the three $\alpha$-carbons and thus is normal to $\mathbf{u}_1$. For the third axis that completes the reference system, $\mathbf{u}_3$, two different options exist, corresponding to the two possible orientations of the cross-product between the first two axes. To avoid problems in the definition, we always choose the solution that makes the reference coordinate system right handed. Also, the vectors between $\alpha$-carbons are normalized before being used in the construction of the reference axes to avoid distortions in the definition due to fluctuations around the average values of the distance between contiguous $\alpha$-carbons. Then, the mathematical expression of the reference system is

$$\boldsymbol{\rho}_{i,i+1} = \frac{\mathbf{r}_{i,i+1}}{|\mathbf{r}_{i,i+1}|} \qquad \boldsymbol{\rho}_{i,i-1} = \frac{\mathbf{r}_{i,i-1}}{|\mathbf{r}_{i,i-1}|} \tag{1}$$

$$\mathbf{u}_1 = \frac{\boldsymbol{\rho}_{i,i+1} \times \boldsymbol{\rho}_{i,i-1}}{|\boldsymbol{\rho}_{i,i+1} \times \boldsymbol{\rho}_{i,i-1}|} \tag{2}$$

$$\mathbf{u}_2 = -\frac{\boldsymbol{\rho}_{i,i+1} + \boldsymbol{\rho}_{i,i-1}}{|\boldsymbol{\rho}_{i,i+1} + \boldsymbol{\rho}_{i,i-1}|} \tag{3}$$

$$\mathbf{u}_3 = \mathbf{u}_1 \times \mathbf{u}_2 \tag{4}$$

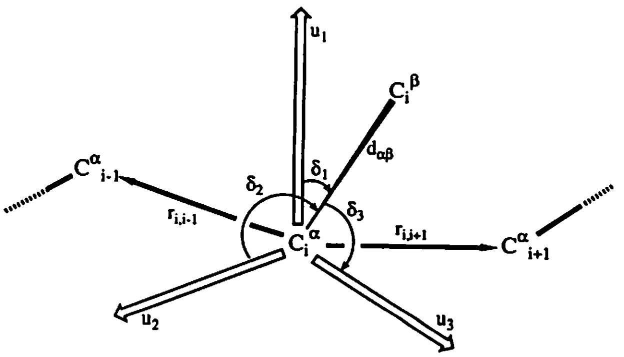


**Figure 1.** Schematic representation of the reference coordinate system that fixes the position of the $\beta$-carbon with respect to the $\alpha$-carbon trace.

and depends exclusively on the coordinates of the two $\alpha$-carbons contiguous to the $i$th $\alpha$-carbon (located in the residue whose reconstruction is being attempted). Thus, this step of the procedure is valid for any residue in the protein with the exception of those situated at the ends of the chain, whose $\beta$-carbon reference system remains undetermined. As shown below, however, a partial reconstruction of the terminal residues can also be handled.

As depicted in Figure 1, the position of $C_i^\beta$ with respect to $C_i^\alpha$ is completely determined by the distance of the chemical bond $C^\alpha$—$C^\beta$, $d_{\alpha\beta}$, and the three director cosines of the angles $\delta_1$, $\delta_2$, and $\delta_3$ this bond forms with the axes of the reference system previously defined. The distance of the chemical bond is almost constant for every residue so if the values of the director cosines also lie in a very narrow range then the localization of the $\beta$-carbon position is straightforward.

To check if this conjecture is true, we scanned a subset of well-refined crystallographic structures in the Protein Data Bank (PDB), and for every internal residue (those not situated at the ends) the construction of the reference system was accomplished and the values of the distance $d_{\alpha\beta}$ and cosine directors $\cos\delta_1$, $\cos\delta_2$, and $\cos\delta_3$ extracted. To examine the consistency of these results, both the average and the rms deviation of these quantities were calculated. An additional test for uniqueness of the values obtained, in the case of the cosine directors, is the proximity of the average values to the geometrical constraint $\cos^2\delta_1 + \cos^2\delta_2 + \cos^2\delta_3 = 1$.

When no distinction is made among different residues, both the deviations in the bond distances and in the director cosines are too large for the average values to be considered acceptable parameters. On the other hand, if an amino-acid-dependent average is computed, deviations are substantially reduced. However, the final mean values of the director cosines deviate from the constraint previously defined (the sum of the squares of the cosines drops in some cases as low as 0.8). This is not a completely unexpected result. One can consider that as the virtual bond angle formed by vectors $\mathbf{r}_{i,i+1}$ and $\mathbf{r}_{i,i-1}$ varies the position of the $C_i^\beta$ will be slightly different to accomodate the new situation. Therefore, our final computation of these geometrical values was done both in an amino-acid-dependent and a distance $d_{\alpha(i-1)-\alpha(i+1)}$-dependent basis (this distance is the separation between $C_{i-1}^\alpha$ and $C_{i+1}^\alpha$). Subject to these conditions, the sum of the squares of the average cosine directors is always larger than 0.93, while the rms deviations are usually less than 1% of the average values. These average values are collected in Table I. They have been normalized so that the sum of their square values equals unity, and thus no further

**Table I.** Distances and director cosines for positioning $\beta$-carbons from the $C^\alpha$ trace.

| | | | $d_{\alpha(i-1)-\alpha(i+1)}$/Å | | | | | |
|---|---|---|---|---|---|---|---|---|
| Residue | $d_{\alpha\beta}$/Å | | (4.5–5.1) | (5.1–5.6) | (5.6–6.1) | (6.1–6.6) | (6.6–7.0) | (7.0–7.4) |
| Gly | 1.070 | $\cos\delta_1$ | 0.3757 | 0.5160 | 0.5273 | 0.5169 | 0.6122 | 0.5606 |
| | | $\cos\delta_2$ | 0.4164 | 0.1553 | 0.1718 | −0.09351 | −0.07438 | −0.01542 |
| | | $\cos\delta_3$ | 0.8279 | 0.8422 | 0.8321 | 0.8509 | 0.7872 | 0.8280 |
| Ala | 1.530 | $\cos\delta_1$ | 0.6971 | 0.6449 | 0.7207 | 0.8657 | 0.8755 | 0.8255 |
| | | $\cos\delta_2$ | −0.1275 | −0.1556 | −0.08253 | −0.1034 | −0.09963 | −0.03687 |
| | | $\cos\delta_3$ | 0.7056 | 0.7482 | 0.6884 | 0.4897 | 0.4728 | 0.5676 |
| Ser | 1.530 | $\cos\delta_1$ | 0.6776 | 0.6829 | 0.7623 | 0.8512 | 0.8694 | 0.8205 |
| | | $\cos\delta_2$ | −0.1096 | −0.1178 | −0.03786 | −0.09848 | −0.09478 | −0.06116 |
| | | $\cos\delta_3$ | 0.7272 | 0.7210 | 0.6461 | 0.5156 | 0.4849 | 0.5684 |
| Cys | 1.528 | $\cos\delta_1$ | 0. | 0.6452 | 0.7445 | 0.8337 | 0.8771 | 0.8271 |
| | | $\cos\delta_2$ | 0. | −0.1641 | −0.07196 | −0.02966 | −0.1084 | −0.05292 |
| | | $\cos\delta_3$ | 0. | 0.7462 | 0.6637 | 0.5515 | 0.4680 | 0.5595 |
| Val | 1.540 | $\cos\delta_1$ | 0. | 0.6800 | 0.8093 | 0.9112 | 0.9159 | 0.8713 |
| | | $\cos\delta_2$ | 0. | −0.1573 | −0.03760 | −0.05248 | −0.03961 | −0.07040 |
| | | $\cos\delta_3$ | 0. | 0.7162 | 0.5861 | 0.4086 | 0.3994 | 0.4856 |
| Thr | 1.560 | $\cos\delta_1$ | 0.6570 | 0.6942 | 0.7684 | 0.8893 | 0.8974 | 0.8382 |
| | | $\cos\delta_2$ | −0.1423 | −0.1121 | −0.03911 | −0.08801 | −0.09638 | −0.09259 |
| | | $\cos\delta_3$ | 0.7403 | 0.7110 | 0.6388 | 0.4488 | 0.4306 | 0.5375 |
| Ile | 1.554 | $\cos\delta_1$ | 0. | 0.6653 | 0.7822 | 0.9158 | 0.9237 | 0.8805 |
| | | $\cos\delta_2$ | 0. | −0.1784 | −0.07882 | −0.03813 | −0.04544 | −0.03661 |
| | | $\cos\delta_3$ | 0. | 0.7250 | 0.6181 | 0.3998 | 0.3803 | 0.4727 |
| Pro trans | 1.527 | $\cos\delta_1$ | 0.5695 | 0.6266 | 0.7396 | 0.8376 | 0.8279 | 0.5471 |
| | | $\cos\delta_2$ | −0.1088 | −0.08485 | −0.01203 | −0.05556 | −0.1142 | −0.2859 |
| | | $\cos\delta_3$ | 0.8148 | 0.7747 | 0.6729 | 0.5435 | 0.5491 | 0.7867 |

**Table I.** (continued)

| Residue | $d_{\alpha\beta}$/Å | | $d_{\alpha(i-1)-\alpha(i+1)}$/Å | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | (4.5–5.1) | (5.1–5.6) | (5.6–6.1) | (6.1–6.6) | (6.6–7.0) | (7.0–7.4) |
| Pro cis | 1.536 | cos $\delta_1$ | 0.9128 | 0.9241 | 0.8906 | 0. | 0. | 0. |
| | | cos $\delta_2$ | −0.3725 | −0.3791 | −0.4545 | 0. | 0. | 0. |
| | | cos $\delta_3$ | 0.1675 | 0.04790 | −0.01835 | 0. | 0. | 0. |
| Met | 1.528 | cos $\delta_1$ | 0. | 0.6901 | 0.7643 | 0.9168 | 0.8829 | 0.8437 |
| | | cos $\delta_2$ | 0. | −0.1462 | −0.05222 | −0.04566 | −0.1008 | −0.05434 |
| | | cos $\delta_3$ | 0. | 0.7087 | 0.6427 | 0.3966 | 0.4586 | 0.5340 |
| Asp | 1.533 | cos $\delta_1$ | 0.6749 | 0.7044 | 0.8113 | 0.8872 | 0.8974 | 0.7365 |
| | | cos $\delta_2$ | −0.1616 | −0.09610 | −0.01579 | −0.009180 | −0.07950 | −0.06435 |
| | | cos $\delta_3$ | 0.7199 | 0.7033 | 0.5844 | 0.4614 | 0.4339 | 0.6733 |
| Asn | 1.534 | cos $\delta_1$ | 0. | 0.6944 | 0.8009 | 0.8859 | 0.8800 | 0.8247 |
| | | cos $\delta_2$ | 0. | 0.03277 | 0.03813 | −0.007092 | −0.06561 | −0.07913 |
| | | cos $\delta_3$ | 0. | 0.7188 | 0.5976 | 0.4638 | 0.4704 | 0.5600 |
| Leu | 1.536 | cos $\delta_1$ | 0.6660 | 0.6573 | 0.7878 | 0.8931 | 0.8816 | 0.8602 |
| | | cos $\delta_2$ | 0.08749 | −0.1408 | −0.04661 | −0.04922 | −0.08487 | −0.05791 |
| | | cos $\delta_3$ | 0.7408 | 0.7404 | 0.6142 | 0.4472 | 0.4643 | 0.5066 |
| Lys | 1.528 | cos $\delta_1$ | 0. | 0.6489 | 0.7417 | 0.8756 | 0.8854 | 0.8482 |
| | | cos $\delta_2$ | 0. | −0.1717 | −0.07902 | −0.08511 | −0.08652 | −0.1117 |
| | | cos $\delta_3$ | 0. | 0.7412 | 0.6660 | 0.4756 | 0.4566 | 0.5179 |
| Glu | 1.531 | cos $\delta_1$ | 0.6119 | 0.6460 | 0.7308 | 0.8769 | 0.8893 | 0.8381 |
| | | cos $\delta_2$ | −0.2138 | −0.1882 | −0.09939 | −0.08536 | −0.09156 | −0.07577 |
| | | cos $\delta_3$ | 0.7615 | 0.7398 | 0.6753 | 0.4730 | 0.4480 | 0.5402 |
| Gln | 1.529 | cos $\delta_1$ | 0.6056 | 0.6625 | 0.7377 | 0.8826 | 0.8952 | 0.8543 |
| | | cos $\delta_2$ | −0.1582 | −0.1519 | −0.06338 | −0.07839 | −0.1010 | −0.07040 |
| | | cos $\delta_3$ | 0.7799 | 0.7335 | 0.6721 | 0.4636 | 0.4341 | 0.5150 |
| Arg | 1.532 | cos $\delta_1$ | 0.6541 | 0.6810 | 0.7404 | 0.8952 | 0.8878 | 0.8209 |
| | | cos $\delta_2$ | −0.2221 | −0.1514 | −0.06434 | −0.07840 | −0.09276 | −0.07128 |
| | | cos $\delta_3$ | 0.7230 | 0.7164 | 0.6691 | 0.4387 | 0.4509 | 0.5667 |
| His | 1.542 | cos $\delta_1$ | 0. | 0.7121 | 0.7612 | 0.8475 | 0.8739 | 0.8261 |
| | | cos $\delta_2$ | 0. | −0.09347 | −0.05836 | −0.08140 | −0.04959 | −0.07366 |
| | | cos $\delta_3$ | 0. | 0.6958 | 0.6458 | 0.5246 | 0.4835 | 0.5586 |
| Phe | 1.534 | cos $\delta_1$ | 0. | 0.6617 | 0.7826 | 0.9175 | 0.9041 | 0.8593 |
| | | cos $\delta_2$ | 0. | −0.1587 | −0.04809 | −0.02742 | −0.1004 | −0.07982 |
| | | cos $\delta_3$ | 0. | 0.7327 | 0.6207 | 0.3968 | 0.4154 | 0.5051 |
| Tyr | 1.541 | cos $\delta_1$ | 0. | 0.6738 | 0.7864 | 0.8790 | 0.9010 | 0.8602 |
| | | cos $\delta_2$ | 0. | −0.09011 | −0.04526 | −0.03848 | −0.06856 | −0.09126 |
| | | cos $\delta_3$ | 0. | 0.7334 | 0.6160 | 0.4753 | 0.4285 | 0.5017 |
| Trp | 1.534 | cos $\delta_1$ | 0. | 0.7030 | 0.7703 | 0.8776 | 0.9028 | 0.8594 |
| | | cos $\delta_2$ | 0. | −0.1222 | −0.006324 | −0.07555 | −0.1082 | −0.06459 |
| | | cos $\delta_3$ | 0. | 0.7006 | 0.6377 | 0.4734 | 0.4162 | 0.5073 |
| Cyx | 1.530 | cos $\delta_1$ | 0. | 0.6817 | 0.7679 | 0.8665 | 0.8861 | 0.7897 |
| | | cos $\delta_2$ | 0. | −0.1275 | −0.09261 | −0.09657 | −0.1004 | −0.07276 |
| | | cos $\delta_3$ | 0. | 0.7205 | 0.6339 | 0.4898 | 0.4524 | 0.6091 |

See the section on the problem of glycines for explanation of glycine values.

corrections are needed to keep the distance $d_{\alpha\beta}$ at the correct value. The definition of the grid corresponding to the distances $d_{\alpha(i-1)-\alpha(i+1)}$ requires some explanation. Although several possibilities are almost equivalent, we have chosen one consistent with a lattice representation previously employed in our group.[9] While the division seems a bit arbitrary (since it is based in the square of the distance $d_{\alpha(i-1)-\alpha(i+1)}$ and not in the distance itself), it has the advantage that the different bins are small enough to discriminate the dependence of the director cosines on $d_{\alpha(i-1)-\alpha(i+1)}$ but still large enough to allow for a statistical population in every one of them that confidently validates the final averages. As a matter of fact, the bins whose population was not large enough are discarded in the calculation of the averages. This is the origin of the missing values (set equal to zero) in Table I. To avoid possible singularities, the final algorithm is implemented so these empty bins are filled with the value in the adjacent bin for a given amino acid.

Another point in Table I is the separation of the values of proline in the cis and trans conformations. This was done during the analysis of the crystallographic structures based on the value of the distance between residues $C_{i-1}^{\alpha}$ and $C_i^{\alpha}$. While the average distance between two contiguous $\alpha$-carbons is about 3.8 Å for the majority of the residues (with a trans configuration of the peptide bond), the distance in a pair whose second residue is a cis proline drops to about 2.9 Å and so a difference between both conformations, clearly reflected in the other geometrical values, is evident.

It is important to realize that the statistical study of the crystallographic structures undertaken in this work is only a way of looking for average values of a series of geometrical quantities, which are subsequently fixed as parameters prior to any attempt at backbone reconstruction. This fact establishes a fundamental difference between the present approach and the alternative methods mentioned in the introduction that employ analysis of known structures in the reconstruction procedure itself.

## BACKBONE RECONSTRUCTION

Once the position of the $\alpha$- and $\beta$-carbons is known for a given residue, one can try to locate the coordinates of the corresponding carbonyl group and amide nitrogen. For this task, one has the information provided by the chemical bonds that join the different atoms. Thus, the distances $d_{\alpha C}$ and $d_{\alpha N}$, together with the angles $\tau_{N\alpha\beta}$, $\tau_{\beta\alpha C}$, and $\tau_{N\alpha C}$, can be considered known parameters in the formulation of the problem (see Fig. 2 for the definition of the different symbols used and Table II for the values of bond lengths and bond angles we are using and that were also obtained form the PDB analysis). In addition, the L chirality of the amino acids occurring in real proteins constitutes additional information. This information, however, is not enough to determine the position of the backbone atoms (C and N, at this stage) since the distorted pyramid formed by $C_i$, $N_i$, $C_i^{\alpha}$, and $C_i^{\beta}$ has free rotation about the edge defined by $C_i^{\alpha}$ and $C_i^{\beta}$ and satisfies all the bond lengths and bond angles when considered on a single-residue basis.

To solve the problem, one has again to consider the residues adjacent to the central one. It is known that the peptide bond can be assumed to be in a planar conformation, with only small fluctuations around the trans value for the torsional angle $\omega$ (with the exception, already mentioned above, of some prolines appearing in the cis conformation, that still keep the planar condition). In this situation, it is evident that the positions of the backbone atoms N and C are coupled between adjacent residues through the geometry of the peptide bond and with the $\beta$-carbons because of the almost tetrahedral valence of the $\alpha$-carbons. This coupling propagates down the chain so that once the $\alpha$-carbon coordinates are fixed a unique solution exists for the backbone conformation. In principle, it is possible to formulate a system including all the restrictions of bond lengths and bond angles, plus the planarity of the peptide bond, through the whole chain. This would yield a supersystem of nonlinear equations from which the coordinates of all nitrogen and carbonyl carbon atoms in the backbone could be obtained. However, the huge dimension of the system generated this way and its nonlinear character make it desirable to look for an alternative solution to the problem.

Such a solution, in fact, exists. The occurrence of the peptide bond in a planar conformation brings as a consequence the existence of two angles, denoted as $\xi$ and $\eta$ in Figure 2, whose values also remain fixed.[10] $\xi$ is the angle subtended between the $C_i^{\alpha}$—$N_i$ bond and the imaginary line that joins $C_i^{\alpha}$ with

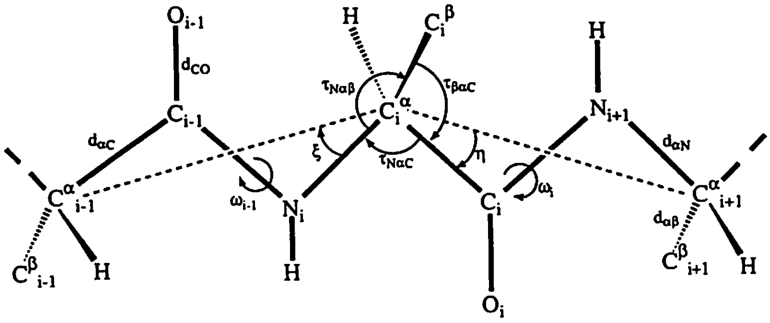


**Figure 2.** Full atom representation of a protein backbone defining the different lengths and angles employed as geometrical constraints.

**Table II.** Bond lengths and angles involved in full atom residue reconstruction.

| Residue | $d_{\alpha N}$/Å | $d_{\alpha C}$/Å | $\tau_{N\alpha C}$ | $\tau_{N\alpha\beta}$ | $\tau_{\beta\alpha C}$ | $\xi$ | $\eta$ |
|---|---|---|---|---|---|---|---|
| Gly | 1.468 | 1.523 | 111.8° | 109.9° | 110.2° | 15.20° | 20.84° |
| Ala | 1.469 | 1.525 | 110.6° | 109.9° | 110.2° | 15.04° | 20.40° |
| Ser | 1.469 | 1.524 | 111.1° | 110.2° | 110.0° | 15.01° | 20.50° |
| Cys | 1.467 | 1.523 | 111.0° | 110.3° | 110.5° | 15.12° | 20.44° |
| Val | 1.472 | 1.530 | 109.4° | 110.8° | 111.9° | 15.05° | 20.60° |
| Thr | 1.471 | 1.525 | 110.4° | 110.9° | 110.9° | 15.13° | 20.38° |
| Ile | 1.472 | 1.528 | 109.5° | 111.1° | 111.6° | 15.05° | 20.66° |
| Pro trans | 1.468 | 1.522 | 111.8° | 104.7° | 111.4° | 15.13° | 21.16° |
| Pro cis | 1.464 | 1.524 | 113.3° | 103.2° | 110.8° | 59.97° | 20.20° |
| Met | 1.469 | 1.527 | 110.9° | 110.9° | 110.6° | 14.97° | 20.64° |
| Asp | 1.468 | 1.527 | 110.9° | 110.7° | 111.1° | 14.90° | 20.45° |
| Asn | 1.472 | 1.527 | 110.6° | 110.1° | 111.4° | 14.89° | 20.39° |
| Leu | 1.469 | 1.527 | 110.4° | 109.4° | 111.2° | 14.99° | 20.49° |
| Lys | 1.469 | 1.524 | 110.7° | 109.9° | 109.5° | 14.83° | 20.47° |
| Glu | 1.468 | 1.522 | 111.3° | 110.9° | 109.2° | 15.10° | 20.63° |
| Gln | 1.469 | 1.526 | 110.9° | 110.7° | 110.4° | 14.89° | 20.65° |
| Arg | 1.473 | 1.523 | 110.5° | 110.9° | 109.9° | 15.19° | 20.51° |
| His | 1.470 | 1.523 | 110.7° | 110.9° | 110.1° | 15.26° | 20.52° |
| Phe | 1.470 | 1.528 | 110.3° | 111.1° | 110.8° | 14.99° | 21.43° |
| Tyr | 1.469 | 1.525 | 110.9° | 110.3° | 110.3° | 15.01° | 21.09° |
| Trp | 1.472 | 1.527 | 110.5° | 110.8° | 110.5° | 15.12° | 20.44° |
| Cyx | 1.471 | 1.527 | 110.5° | 110.1° | 109.7° | 15.31° | 20.56° |

$C^\alpha_{i-1}$, while $\eta$ is the angle formed by the $C^\alpha_i$—$C_i$ bond and the imaginary line that joins $C^\alpha_i$ with $C^\alpha_{i+1}$. Thus, these two angles represent additional geometrical requirements that every single-residue conformation has to fulfill. Even more, they are enough to fully determine the orientation of the pyramid with respect to the $\alpha$-carbon backbone and therefore allow to position the coordinates of $N_i$ and $C_i$.

To completely clarify the different geometrical restrictions available, let us formulate them in a mathematical form. If we express the position of the different atoms as Cartesian coordinates with respect to the central $\alpha$-carbon, the equations that must be satisfied are:

- bond lengths

$$(x_i^N)^2 + (y_i^N)^2 + (z_i^N)^2 - d_{\alpha N}^2 = 0 \qquad (5)$$

$$(x_i^C)^2 + (y_i^C)^2 + (z_i^C)^2 - d_{\alpha C}^2 = 0. \qquad (6)$$

- bond angles

$$x_i^N x_i^\beta + y_i^N y_i^\beta + z_i^N z_i^\beta - d_{\alpha N} d_{\alpha\beta} \cos \tau_{N\alpha\beta} = 0 \qquad (7)$$

$$x_i^N x_i^C + y_i^N y_i^C + z_i^N z_i^C - d_{\alpha N} d_{\alpha C} \cos \tau_{N\alpha C} = 0 \qquad (8)$$

$$x_i^C x_i^\beta + y_i^C y_i^\beta + z_i^C z_i^\beta - d_{\alpha C} d_{\alpha\beta} \cos \tau_{\beta\alpha C} = 0 \qquad (9)$$

We should not forget that since the position of the $\beta$-carbon was previously determined the coordinates $(x_i^\beta, y_i^\beta, z_i^\beta)$ have known values in this system.

- *$\alpha$-carbon chirality.* There are several ways to express the appropriate chirality of the $\alpha$-carbon. We have chosen one based on the expression of the scalar triple product of the vectors that, with origin in $C^\alpha_i$, point towards $N_i$, $C^\beta_i$, and $C_i$. To satisfy the $L$ chirality, it is enough that

$$Tr(i) = \mathbf{r}_{\alpha N} \cdot (\mathbf{r}_{\alpha\beta} \times \mathbf{r}_{\alpha C}) = \begin{vmatrix} x_i^N & y_i^N & z_i^N \\ x_i^\beta & y_i^\beta & z_i^\beta \\ x_i^C & y_i^C & z_i^C \end{vmatrix} < 0. \qquad (10)$$

This inequality, however, can be developed further by expressing the value of $Tr(i)$ in terms of bonds lengths and bond angles. By doing so, one can arrive at the final equation

$$Tr(i) = -d_{\alpha N} d_{\alpha\beta} d_{\alpha C} (1 - \cos^2 \tau_{N\alpha C} - \cos^2 \tau_{N\alpha\beta} - \cos^2 \tau_{\beta\alpha C} + 2 \cos \tau_{N\alpha C} \cos \tau_{N\alpha\beta} \cos \tau_{\beta\alpha C})^{1/2} \qquad (11)$$

- Angles $\eta$ and $\xi$

$$x_i^N x_{i-1}^\alpha + y_i^N y_{i-1}^\alpha + z_i^N z_{i-1}^\alpha - d_{\alpha N} d_{\alpha_i \alpha_{i-1}} \cos \xi = 0 \qquad (12)$$

$$x_i^C x_{i+1}^\alpha + y_i^C y_{i+1}^\alpha + z_i^C z_{i+1}^\alpha - d_{\alpha C} d_{\alpha_i \alpha_{i+1}} \cos \eta = 0. \qquad (13)$$

In these two equations, $d_{\alpha_i\alpha_{i-1}}$ and $d_{\alpha_i\alpha_{i+1}}$ are the distances from the $i$th $\alpha$-carbon to the previous and following $\alpha$-carbons, respectively. Also, the coordinates $(x^\alpha_{i\pm1}, y^\alpha_{i\pm1}, z^\alpha_{i\pm1})$ of these two $\alpha$-carbons are defined with respect to $C^\alpha_i$, i.e., $x^\alpha_{i\pm1} = X^\alpha_{i\pm1} - X^\alpha_i$, where the capital case variables correspond to the absolute coordinates, whose origin is an external reference frame (e.g., they are the coordinates of $\alpha$-carbons as provided by the Protein Data Bank or any other experimental source).

Equations (5)–(9) and (11)–(13) are not independent, as it is obvious when considering that, at this moment, we have formulated eight equations for only a set of six unknowns (the Cartesian coordinates of C and N with respect to the $\alpha$-carbon). It is evident, e.g., that the equations that refer to bond lengths and bond angles are partially redundant with the equation of the triple product. Therefore, it would be possible in principle to find a subset of independent equations that would determine a completely defined (although still nonlinear) system whose analytical or numerical solution would provide the desired coordinates. This approach, however, brings some additional complications. Although we are assuming that bond lengths and bond angles have definite values for a given residue, this is not what one finds in real experimental data. Of course, some large deviations can be attributed to an incorrect refinement procedure of the crystal structures, but even when this task is undertaken with extreme care small fluctuations around the average values of the geometrical constraints still arise and are, as a matter of fact, necessary to fit the structure to experiment. Nevertheless, when one faces a given structure it is impossible to know whether or not the observed distortions represent physical reality. In these conditions, it seems safer to keep as much experimental information as possible and try to find a compromise solution that satisfies, as best as possible, all the geometrical requirements. The problems found by Purisima and Scheraga in their analytical method[8] also support this viewpoint. That is why we have kept all the above equations in our analytical method. Obviously, the mathematical problem is transformed from the solution of a set of nonlinear equations to the simultaneous minimization of a larger set of nonlinear equations. Even with this in mind, there are several equations that can be considered more "reliable" than others. For example, the values of the angles $\xi$ and $\eta$ present the largest fluctuations of all the average values obtained from the crystallographic structures. The reason is that their constancy is based upon the assumption of planar and rigid peptide bonds and ignore the fluctuations that occur in real structures. In addition, eqs. (7), (9), and (11) include the position of the $\beta$-carbons, which can be also slightly affected by the use of average constant values in their determination. On the other hand, eqs. (5) and (6) include only bond lengths whose fluctuations in well-refined structures are very small. Equation (8) also includes the bond angle $\tau_{N\alpha C}$, whose value is also quite well established on its own since it is completely independent (at least from the geometrical point of view) of the $\beta$-carbon position. Thus, we have chosen to minimize the sum of the squares of eqs. (7) and (9)–(13), keeping eqs. (5), (6), and (8) as mathematical constraints. This is equivalent to moving the rigid set of three atoms N—C$^\alpha$—C, with the correct distances and internal angle, around the vertex situated in the C$^\alpha$ position, until the five equations included in the minimization set are satisfied as closely as possible.

The numerical solution of the problem formulated above has been accomplished using the DNCONG routine in the IMSL library.[11] This subroutine is based on the use of a successive quadratic programming method.[12] The algorithm requires the user to supply the gradient of the minimization function and of the mathematical constraints with respect to the unknowns included in the system. This is, in fact, a quite straightforward calculation and does not deserve any further comment.

By using this algorithm, we get a solution for the relative coordinates ($x^N$, $y^N$, $z^N$) and ($x^C$, $y^C$, $z^C$) that usually satisfies all geometrical equations reasonably well. The transformation of these relative coordinates to absolute values involves only adding the $C_i^\alpha$ coordinates. Only in a few cases, usually associated with glycines (whose particular consideration we shall discuss immediately below) and prolines in the cis conformation, is the quality of the minimization not so good. In subsequent sections, we discuss possible ways of refining the resulting coordinates.

## THE PROBLEM OF GLYCINES

As we have seen, the basis of the backbone reconstruction lies in the previous positioning of the $\beta$-carbons. Thus, when one glycine appears in the primary sequence of the protein a gap would result in the sequence of backbone atoms. As we shall see when talking about the residues situated at the ends of the chain, this does not pose a real problem when a single isolated glycine appears in the sequence. However, two or more contiguous glycines would represent a nonresolvable problem. To avoid this complication, we used a small (although quite valid) trick, consistent in defining the L hydrogen of the glycines, and use it as the equivalent to the $\beta$-carbon (although with the correct bond distance) in all subsequent calculations. To do that, we included a modification in the program that scans the experimental structures looking for the cosine directors of the $\beta$-carbons. When a glycine appears, instead of skipping it, the distorted tetrahedron centered in the $\alpha$-carbon and with two vertices in the positions of nitrogen and carbon is constructed. After that, the vertex corresponding to the L configuration is chosen as the direction of the hydrogen, whose director cosines are then determined from the usual reference system. Of course, the method involves a certain ambiguity in the definition of the tetrahedral positions, and so the full atom reconstruction of glycines is, on average, worse than for the other amino acids. However, the procedure works rather well in most of the cases. The director cosines resultant from this analysis are also included in Table I.

## TERMINAL RESIDUES

For the two residues situated at the ends of the primary sequence, the definition of an internal reference system that allows positioning of the $\beta$-carbons (or the L hydrogen for glycines) is not possible, and this precludes the possibility of locating the backbone atoms for these two residues via the previously described process. Even though there is no way of solving this problem for the $\beta$-carbon or the terminal backbone atom (the N in the first residue and the C═0 in the last residue), it is possible to locate the coordinates of the atoms involved in the peptide bond once the plane in which this lies is defined by the next (or the previous) residue, whose backbone atoms must have been already determined.

Figure 3 shows the geometrical constraints that must now be satisfied. We have suppressed the $\beta$-carbons since the position of one of them is unknown in our reconstruction procedure and the other is unnecessary. The available information in this case consists of the distance of the desired atom (the carbon in the N-terminus or the nitrogen in the C-terminus) to the corresponding $\alpha$-carbon, the distance and angles of the peptide bond, its planarity (the plane being defined by the position of the $\alpha$-carbons that delineate it and the atom—N or C—whose position is already known), and the value of the angle $\xi$ or $\eta$, as defined in Figures 2 and 3.

Then, the equations that need to be satisfied now are:

- *N-terminus.* The unknowns are the three coordinates $(x_1^C, y_1^C, z_1^C)$ and it is assumed that the coordinates $(X_2^N, Y_2^N, Z_2^N)$ are already known. Since the center of the reference system is taken in $C_1^\alpha$, these last coordinates (which we have expressed in upper case symbols, indicating its absolute meaning) have to be recomputed with respect to the chosen origin.

$$(x_1^C)^2 + (y_1^C)^2 + (z_1^C)^2 - d_{\alpha C}^2 = 0 \quad (14)$$

$$[x_1^C - (X_2^N - X_1^\alpha)]^2 + [y_1^C - (Y_2^N - Y_1^\alpha)]^2 + [z_1^C - (Z_2^N - Z_1^\alpha)]^2 - d_{CN}^2 = 0 \quad (15)$$

$$x_1^C[x_1^C - (X_2^N - X_1^\alpha)] + y_1^C[y_1^C - (Y_2^N - Y_1^\alpha)] + z_1^C[z_1^C - (Z_2^N - Z_1^\alpha)] - d_{CN}d_{\alpha C} \cos \tau_{\alpha CN} = 0 \quad (16)$$

$$\mathbf{r}_{\alpha_1 C_1} \cdot (\mathbf{r}_{\alpha_1 N_2} \times \mathbf{r}_{\alpha_1 \alpha_2}) = \begin{vmatrix} x_1^C & y_1^C & z_1^C \\ X_2^N - X_1^\alpha & Y_2^N - Y_1^\alpha & Z_2^N - Z_1^\alpha \\ X_2^\alpha - X_1^\alpha & Y_2^\alpha - Y_1^\alpha & Z_2^\alpha - Z_1^\alpha \end{vmatrix} = 0 \quad (17)$$

$$x_1^C(X_2^\alpha - X_1^\alpha) + y_1^C(Y_2^\alpha - Y_1^\alpha) + z_1^C(Z_2^\alpha - Z_1^\alpha) - d_{\alpha C}d_{\alpha_1\alpha_2} \cos \eta = 0. \quad (18)$$

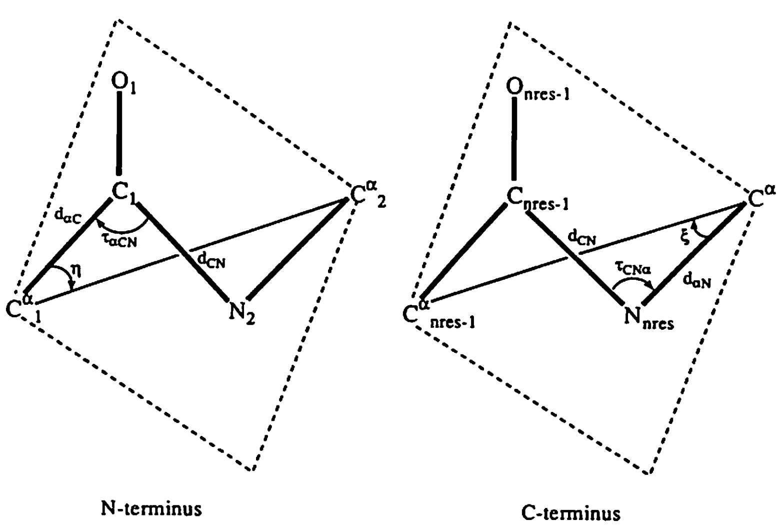
- *C-terminus.* In this case, the three unknowns are the coordinates $(x_{nres}^N, y_{nres}^n, z_{nres}^N)$, with *nres* being the number of residues of the protein, and the same considerations have to be made about the origin of the reference system, centered now at $C_{nres}^\alpha$.

$$(x_{nres}^N)^2 + (y_{nres}^N)^2 + (z_{nres}^N)^2 - d_{\alpha N}^2 = 0 \quad (19)$$

$$[x_{nres}^N - (X_{nres-1}^C - X_{nres}^\alpha)]^2 + [y_{nres}^N - (Y_{nres-1}^C - Y_{nres}^\alpha)]^2 + [z_{nres}^C - (Z_{nres-1}^N - Z_{nres}^\alpha)]^2 - d_{CN}^2 = 0 \quad (20)$$

$$x_{nres}^N[x_{nres}^N - (X_{nres-1}^C - X_{nres-1}^\alpha)] + y_{nres}^N\,[y_{nres}^N - (Y_{nres-1}^C - Y_{nres-1}^\alpha)] + z_{nres}^N\,[z_{nres}^N - (Z_{nres-1}^C - Z_{nres-1}^\alpha)] - d_{CN}d_{\alpha N} \cos \tau_{CN\alpha} = 0 \quad (21)$$

$$\mathbf{r}_{\alpha_{nres}N_{nres}} \cdot (\mathbf{r}_{\alpha_{nres}C_{nres-1}} \times \mathbf{r}_{\alpha_{nres}\alpha_{nres-1}}) = \begin{vmatrix} x_{nres}^N & y_{nres}^N & z_{nres}^N \\ X_{nres-1}^C - X_{nres}^\alpha & Y_{nres-1}^C - Y_{nres}^\alpha & Z_{nres-1}^C - Z_{nres}^\alpha \\ X_{nres-1}^\alpha - X_{nres}^\alpha & Y_{nres-1}^\alpha - Y_{nres}^\alpha & Z_{nres-1}^\alpha - Z_{nres}^\alpha \end{vmatrix} = 0 \quad (22)$$



N-terminus

C-terminus

**Figure 3.** Atoms and geometrical values involved in the coordinate determination of $C_1$ and $N_{nres}$. The broken lines sketch the peptide plane.

$$x^N_{nres}(X^\alpha_{nres-1} - X^\alpha_{nres}) + y^N_{nres}(Y^\alpha_{nres-1} - Y^\alpha_{nres}) + z^N_{nres}(Z^\alpha_{nres-1} - Z^\alpha_{nres}) - d_{\alpha C} d_{\alpha_{nres}\alpha_{nres-1}} \cos \xi = 0. \quad (23)$$

Again, we keep more equations than unknowns. This is especially important in these conditions since three of the five equations in every set depend on some previously determined coordinates and can therefore include some error. In this case, we have considered all the formulated equations as part of the minimization function without any of them acting as a mathematical constraint. The values of $d_{CN}$, $\tau_{CN\alpha}$, and $\tau_{\alpha CN}$ included in these equations are listed in Table III and, as usual, correspond to the average of the values found in the PDB structures.

## POSITION OF THE CARBONYL OXYGENS

Once all the consecutive atoms of the backbone have been positioned, the coordinates of the oxygens corresponding to the carbonyl groups should be, in principle, a trivial calculation. Unfortunately, this does not hold in reality and serves to bring us to some of the limitations in the model. These problems are related to the "isolated" character of the reconstruction procedure, in which every residue is considered independently (along with information provided by the $\alpha$-carbon positions of the adjacent residues). This means that, with the exception of the end residues, the planarity of the peptide bond is not *explicitly* included into the system of equations. We have emphasized the term explicitly because, as stated before, equations involving the angles $\xi$ and $\eta$ are a direct consequence of this planarity. However, both due to the fluctuations found in the crystallographic structures for these angles and numerical errors in the solution of the system of equations, one finds that the final planarity of the peptide bond is not absolutely achieved in all cases. Many times, the fluctuations of the angle $\omega$ that defines the torsional state of the peptide bond (see Fig. 2) are only a few degrees (less than 15°, with a high percentage being less than 10°, as happens in experimental structures). However, a few residues in every reconstructed protein exhibit larger deviations, even up to 60–90°. The reasons for these deviations are not clear at the moment. However, they pose a serious problem when trying to position the oxygen atom. If the vectors joining $C^\alpha_i$—$C_i$ and $N_{i+1}$—$C^\alpha_{i+1}$ are not in the same plane, the determination of the oxygen position is subject to a certain ambiguity.

**Table III.** Bond lengths and bond angles involved in the reconstruction of terminal residues.

| Residue | $d_{CN}$/Å | $\tau_{\alpha CN}$ | $\tau_{CN\alpha}$ |
|---|---|---|---|
| Gly | 1.322 | 115.7° | 121.2° |
| Ala | 1.322 | 116.2° | 121.4° |
| Ser | 1.321 | 115.8° | 121.3° |
| Cys | 1.322 | 115.9° | 121.3° |
| Val | 1.321 | 115.8° | 121.5° |
| Thr | 1.322 | 116.0° | 121.2° |
| Ile | 1.320 | 115.6° | 121.4° |
| Pro trans | 1.321 | 115.8° | 121.8° |
| Pro cis | 1.323 | 116.5° | 124.6° |
| Met | 1.322 | 115.8° | 121.4° |
| Asp | 1.322 | 115.9° | 121.6° |
| Asn | 1.323 | 116.0° | 121.5° |
| Leu | 1.319 | 116.1° | 121.5° |
| Lys | 1.321 | 115.9° | 121.7° |
| Glu | 1.322 | 116.0° | 121.4° |
| Gln | 1.322 | 116.0° | 121.6° |
| Arg | 1.322 | 116.0° | 121.2° |
| His | 1.322 | 116.0° | 121.3° |
| Phe | 1.322 | 115.6° | 121.5° |
| Tyr | 1.323 | 115.6° | 121.5° |
| Trp | 1.326 | 116.0° | 121.5° |
| Cyx | 1.319 | 115.6° | 120.9° |

To avoid this as much as possible, we tried to establish a series of conditions that do not only depend on the previously determined atomic coordinates. In Figure 4, a scheme showing the associated geometrical restrictions is depicted. The distance $d_{\alpha O}$ between an $\alpha$-carbon and the oxygen in the same residue was determined from the crystallographic structures analysis, as was the angle $\tau_{O\alpha\alpha}$ that forms the imaginary lines $C^\alpha_i$—$O_i$ and $C^\alpha_i$—$C^\alpha_{i+1}$. On the other hand, the choice of the plane in which the oxygen atom is going to be positioned (and that constitutes the only additional information required to locate it) has to be determined from the atoms involved in the peptide bond. After a few tests, we found that it was usually slightly better to define the plane from the position of the nitrogen $N_{i+1}$ than from carbonyl carbon $C_i$ since the calculated coordinates of the nitrogen yield an average deviation from the PDB coordinates that is smaller than that for the coordinates of the carbon atom. Obviously, the other two points necessary for defining the plane are the $\alpha$-carbons $C^\alpha_i$ and $C^\alpha_{i+1}$. In these conditions, some cases appear in which the distance $d_{CO}$ or the angles $\tau_{\alpha CO}$ and $\tau_{OCN}$ are distorted with respect to their average values. However, the alternative procedure of including these additional geometrical requirements and trying to find a minimum for the whole set of equations did not in general improve the result. The values of $d_{CO}$ and $\tau_{O\alpha\alpha}$ for the different residues are included in Table IV.

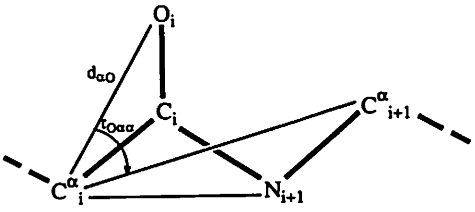


**Figure 4.** Definition of the geometrical constraints employed in the oxygen reconstruction.

**Table IV.** Virtual lengths and angles employed in determination of the oxygen coordinates.

| Residue | $d_{\alpha O}$/Å | $\tau_{O\alpha\alpha}$ |
|---|---|---|
| Gly | 2.401 | 47.38° |
| Ala | 2.401 | 46.97° |
| Ser | 2.396 | 47.15° |
| Cys | 2.397 | 47.07° |
| Val | 2.401 | 47.21° |
| Thr | 2.397 | 47.11° |
| Ile | 2.400 | 42.78° |
| Pro trans | 2.406 | 47.45° |
| Pro cis | 2.405 | 46.52° |
| Met | 2.399 | 47.12° |
| Asp | 2.396 | 47.12° |
| Asn | 2.395 | 47.20° |
| Leu | 2.395 | 47.22° |
| Lys | 2.400 | 46.96° |
| Glu | 2.396 | 47.22° |
| Gln | 2.398 | 47.27° |
| Arg | 2.396 | 47.12° |
| His | 2.398 | 47.09° |
| Phe | 2.399 | 47.89° |
| Tyr | 2.397 | 47.64° |
| Trp | 2.403 | 46.91° |
| Cyx | 2.397 | 47.31° |

## RESULTS AND DISCUSSION

With all the geometrical relations seen so far, we have developed a Fortran algorithm that, having as input information only the coordinates of the $\alpha$-carbons and the nature of the amino acids that comprise the protein (i.e., its primary sequence) ultimately provides the coordinates of all the $\beta$-carbons (with the exception of the glycines and the end residues) and the corresponding positions for the carbon and oxygen in the carbonyl groups and the nitrogen in the amide groups. To check the mathematical consistency of the algorithm, we tested it on some ideal structures built in such a way that they rigorously satisfy the geometrical constraints included in the mathematical equations of our procedure. When the coordinates of the $\alpha$-carbons for these structures are used as input data in our program, the resulting backbone atom coordinates present a negligible deviation with respect to the ideal chain.

Much more important, however, is the ability of the method to handle the nonideal structures that occur in nature. In Table V, we present the resulting coordinate rms deviations for several real proteins (4pti: bovine pancreatic trypsin inhibitor, BPTI; 1pcy; plastocyanin; 2mhr: myohemerythrin; 2lym: lysozyme; 3fxn: flavodoxin; 1tim: triosephosphate isomerase). The $\alpha$-carbon coordinates were extracted from the corresponding files in the Brookhaven Protein Data Bank (with the exception of 1tim[13]) and were chosen to span different possibilities both with respect to the protein size and the elements of secondary structure that constitute the protein. We do not find any dependence of the method on these variables.

It is important to notice the low rms shown by the $\beta$-carbons, whose positions, as we stated before, are the first step in the reconstruction procedure. Almost the same can be said about the nitrogen atoms, and with a slightly less enthusiasm for the carbonyl carbons (the fact that these C atoms present an average deviation larger than nitrogens was the reason we chose the position of the nitrogen as the reference point when determining the coordinates of the oxygens). Nevertheless, when going to the oxygen coordinates the deviations are rather considerable. There are several reasons for this. First, they result from the accumulation of errors involved in the oxygen reconstruction itself (due to the use of statistical average values for the length $d_{\alpha O}$ and the angle $\tau_{O\alpha\alpha}$) and those existing in the position of the nitrogen atom. Second, and more important, the determination of the oxygen coordinates is based upon the assumption of exact planarity of the peptide bond, an assumption that, even ignoring some cases of large deviations, is never completely fulfilled.

To solve this problem, it would be desirable to develop a method that is able to enforce peptide bond planarity and that could be included as a constituent part of the reconstruction procedure itself. This could be accomplished by determining the coordinates of the nitrogen and carbonyl carbon for a given residue inside the chain and then propagating the solution by including in the set of equations to be minimized the condition of peptide bond planarity. This idea, although quite attractive from a con-

**Table V.** Summary of coordinate rms deviations between analytically rebuilt structures and PDB files.

| Protein | No. residues[a] | No. atoms[b] | $C^{\beta}$ | N | C | O | Total |
|---|---|---|---|---|---|---|---|
| 4pti | 58 | 279 | 0.257 | 0.306 | 0.329 | 1.287 | 0.626 |
| 1pcy | 99 | 480 | 0.343 | 0.333 | 0.415 | 1.478 | 0.725 |
| 2mhr | 118 | 580 | 0.294 | 0.371 | 0.421 | 1.452 | 0.711 |
| 2lym | 129 | 628 | 0.289 | 0.356 | 0.437 | 1.552 | 0.756 |
| 3fxn | 138 | 671 | 0.324 | 0.300 | 0.374 | 1.464 | 0.713 |
| 1tim | 249 | 1221 | 0.295 | 0.284 | 0.369 | 1.278 | 0.626 |

rms is expressed in Å.

[a] Number of residues comprising the primary sequence of the protein.

[b] Heavy atoms whose coordinates have been determined ($\alpha$-carbons are not included).

ceptual point of view, is not feasible. As a matter of fact, the algorithm in this case would be identical to that presented by Purisima and Scheraga,[8] with the only formal difference that it is formulated in terms of Cartesian coordinates instead of torsional angles. The fact that all the geometrical equations are included instead of only a minimum set probably would make things work somewhat better, but the method would still strongly depend upon the choice of the residue whose reconstruction is carried out in the first place and whose coordinates act as a seed to propagate the whole reconstruction afterward.

A second possibility would be to include some kind of refinement of the structures after the analytical reconstruction procedure described here has been used. In this sense, we tried the following mechanism. We can assume that the set of $C^\beta$, N, and C for a given residue, together with its $\alpha$-carbon, satisfy the appropriate bond lengths and bond angles. If this is so, we would be able to treat this set of atoms as a rigid body and, by choosing a set of rotations pivoted about the position of the $\alpha$-carbon, slightly modify the orientation of the atoms in every residue with respect to the previous and the following residue with the object of improving the planarity of the two corresponding peptide bonds. Of course, we have to implement an iterative procedure since every movement of one residue affects the two adjacent peptide bonds. To check improvements resulting from this scheme, we added an optimization procedure after the first reconstruction. For every residue, we formulate a set of geometrical relations similar to those employed in the reconstruction of the terminal residues, shown in eqs. (14)–(23). The only difference is that these equations are not formulated now in terms of the Cartesian coordinates of N and C, but include the dependence of these quantities on the three Euler angles that describe the rotation of the central residue with respect to its original position,[14] where the $C^\alpha$ remains fixed in its original coordinates taken from PDB. When the iterative procedure is run trying to minimize the geometrical constraints of the peptide bonds, the rms of the structure begins to be reduced, but after a few iterations it again grows. The method does not always reach convergence, but when it does the final set of coordinates always has a larger deviation than the original structure. There is a reason for this. Since the geometry of the peptide bond is the only requirement imposed on the system during this procedure, there are multiple solutions to the problem because the peptide planes have free rotation around the axis that joins two contiguous $\alpha$-carbons. Of course, the angle $\tau_{N\alpha C}$ couples the different orientations of the peptide planes, but once one of them has been fixed arbitrarily it is possible to find a solution for the others in most of the cases.[8] The question whether these solutions are valid or not depends on the position of the $\beta$-carbons, which follow the rotation movement, and its consistency with the $\alpha$-carbon trace. But, since this consistency is not included in the iterative procedure, the "optimization" continues until one minimum for the peptide bond equations is found that is not going to correspond in general to the correct conformation of the protein backbone.

On the other hand, if the position of the $\beta$-carbons are included in this method (we do this by keeping its coordinates confined to a narrow cone around the original location), the algorithm is not able to produce any real improvement. It is clear that some of the $\beta$-carbons have to be moved more than others to find the correct conformation for the peptide bond, but since the algorithm is not able to recognize which initial estimations of the $C^\beta$ positions are very good and which are not (we remind the reader once more that the information provided to the algorithm consists only in the $C^\alpha$ coordinates) there is not any readily apparent way of designing a purely analytical refinement procedure.

Of course, this difficulty would be readily overcome if any additional experimental information were available. The exact position of a $\beta$-carbon or, even better, of one of the atoms in the backbone, would open up the possibility of very easily using one of the refinement options previously mentioned, propagating the refinement of the coordinates from the point in which they are known to be correct.

Without this information, there is still one procedure that can be used, although the analytical character of the method is then lost, viz., energy minimization. To prepare the backbone, the primary sequence is redefined so that it is only composed of alanines (in all residues whose $\beta$-carbon position has been determined) and glycines (corresponding to the actual glycines of the protein and the terminal residues). By doing this, one can use the default options of some of the standard minimization packages with the certainty that the assignment of charges and hydrogens is going to be correct (with the exception of the end residues since the first N and the last C=O are impossible to locate). In particular, we have used the minimization procedure included in the SYBYL package.[15] All the hydrogen atoms were added to the backbone, as well as the partial charges

**Table VI.** Summary of coordinate rms deviations after energy minimization of the rebuilt backbone.

| Protein | $C^\beta$ | N | C | O | Total |
|---|---|---|---|---|---|
| 4pti | 0.257 | 0.256 | 0.256 | 0.842 | 0.428 |
| 1pcy | 0.350 | 0.337 | 0.349 | 0.843 | 0.464 |
| 2mhr | 0.363 | 0.376 | 0.340 | 0.913 | 0.495 |
| 2lym | 0.317 | 0.228 | 0.261 | 0.854 | 0.438 |
| 3fxn | 0.343 | 0.266 | 0.303 | 0.965 | 0.495 |
| 1tim | 0.294 | 0.213 | 0.245 | 0.759 | 0.393 |

The reference coordinates are still those corresponding to the PDB files. rms is expressed in Å.

corresponding to the resulting structure. The $\alpha$-carbons were defined as aggregates so their positions remain fixed during the process, and a standard energy minimization was run until convergence. Table VI shows the deviations found in the resulting structures compared with the reference crystallographic coordinates. Also, in Figure 5 we show a detailed picture of the distribution of deviations along the primary sequence for flavodoxin (3fxn) both for the structure resulting from the purely analytical reconstruction (left column) and for the energy-minimized structure (right column).

The first remarkable result is that the average deviation of the $\beta$-carbon does not improve at all, but it goes in general to larger values. This does not mean that they remain fixed during the minimization, as

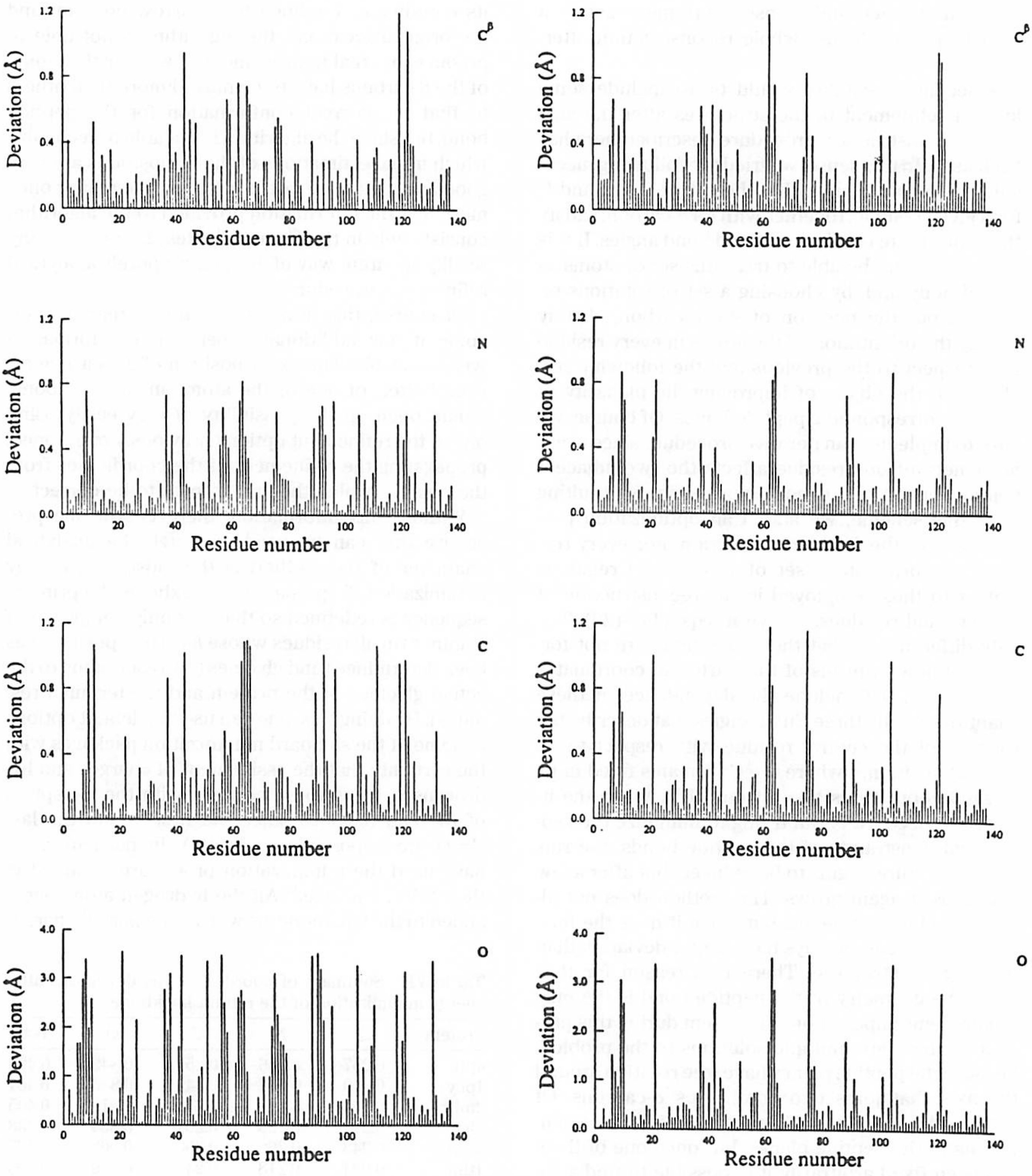


**Figure 5.** Coordinate rms deviations for flavodoxin. Plots at the left column correspond to analytical reconstruction, while the right column corresponds to the structure after energy minimization.

can be observed from the different locations of the dominant deviations in $C^\beta$ shown in Figure 5. On the other hand, the positions of carbons and nitrogens slightly improve, and it is for the oxygens where the positive consequences of the refinement method result are clearly apparent. From this information, it is evident that the main effect of energy minimization is equivalent to the refinement of the peptide bond planarity we tried to implement previously. It looks, however, that while the pure geometrical constraints were not able to find the correct solution for the system, the introduction of the global energy as the main criterion is able to yield a good structure. Of course, it is not exactly equal to the native one. The existence of multiple local minima in which the minimization iterations can be trapped, and more importantly the suppression of the side chains, may considerably modify the energy surface and, at least in some regions, make the system evolve toward a close but different (nonnative) conformation. Also, the choice of the force field has its influence. The structures included in the PDB are not usually the simple result of the crystallographic experiment but are themselves refined through the use of different methods that optimize the resulting geometry, energy, and similarity between the estimated and experimental structure factors. This means that different refinement procedures can yield slightly different sets of coordinates. To check this fact, we ran the energy minimization procedure for the PDB backbone of flavodoxin exactly in the same conditions used for the analytically rebuilt structures (i.e., changing the primary sequence and even suppressing the terminal atoms that are not located in the rebuilt chains). When the minimization reaches convergence, the total rms deviation for the coordinates, with respect to the original PDB file is 0.21Å, half the deviation of the rebuilt structure after the same minimization. Thus, when one is moving on a distance scale of a few tenths of an Å every detail is important, but the narrow distinction between the quality of the different structures is difficult to establish accurately since this region is far beyond the capabilities of the available experimental data themselves.

## CONCLUSIONS

In this article, we presented a method that allows for the analytical reconstruction of a protein backbone with the only information being the primary sequence and a reasonable set of $\alpha$-carbon coordinates. The method is able to give structures whose coordinates deviate on average about 0.7Å from the actual PDB coordinates. This deviation includes the positions of $\beta$-carbons, carbonyl groups, and amide nitrogens. If oxygen coordinates are excluded from these calculations, then the rms drops to half this value. The advantage of this method with respect to those based on PDB scanning or the propagation of a construction/optimization scheme is that the required time is much shorter (the algorithm takes only a few seconds on a Convex C-240 in comparison with several hours of computer time employed by some of the fragment-fitting methods) and in some occasions it provides results at the same level of accuracy. Even the inclusion of the energy minimization keeps the total computational time at a quite reasonable level (on the order of half an hour for the largest protein considered here). Specifically, our analytical results for flavodoxin are comparable to those provided by Correa's method[6] after the backbone reconstruction (which, in his case, already includes a number of energy minimizations). Also, the results for the real structure of BPTI (4pti) are considerably better than those provided by Purisima and Scheraga,[8] although the large rms found by them is not due to the position of the atoms we are considering here but results from the discrepancy between the ideal $C^\alpha$ trace and the real one.

Unfortunately, there is no way of extending this analytical method to the side chains. The inclusion of the torsional angles $\chi$ quickly increases the number of degrees of freedom, and effects such as the packing of the side chains or more specific interactions that are not uniquely related to the $\alpha$-carbon trace preclude any possibility of using geometrical criteria alone for trying to find even an approximate set of coordinates for the side-chain atoms. It is true that, for some amino acids, the spectrum of orientations with respect to the backbone is not continuous.[16] Instead, some kind of "rotamer library" can be found, where a limited set of possible conformations can be defined by a series of director cosines, as done here for the $\beta$-carbons. Even when this set is small enough (reduced to three of four possibilities), there is not information enough to know which rotamer corresponds to a given residue in every situation. Thus, the only way to proceed would be to make an initial guess and run expensive optimizations as in the other reconstruction methods previously proposed.

This fact, however, does not diminish the importance of our algorithm. The results for the backbone alone have proven to be comparable, and sometimes even better, than those obtained from other methods. In addition, the flexibility in the solution of the geometrical constraints, together with the independence of the procedure with respect to real structures, makes it an ideal tool for use in theoretical representations of a protein projected onto a lattice or any other kind of simplified model, and therefore it can significantly contribute to the improvement of these models once the simplified representation has been exploited to its maximum extent.

This work was supported in part by NIH grant No. GM-37408. A.R. also acknowledges a M.E.C./Fulbright Scholarship from the U.S.–Spanish Joint Committee for Cultural and Educational Cooperation.

## References

1. F.C. Bernstein, T.F. Koetzle, E.J.B. Williams, E.F. Meyer Jr., O. Kennard, T. Shimanouchi and M. Tasumi, *J. Mol. Biol.* **112,** 535 (1977).
2. J.A. McCammon and S.C. Harvey, *Dynamics of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK, 1987.
3. T.A. Jones and S. Thirup, *EMBO J.*, **5,** 819 (1986).
4. L.S. Reid and J.M. Thornton, *Proteins*, **5,** 170 (1989).
5. L. Holm and C. Sander, *J. Mol. Biol.*, **218,** 183 (1991).
6. P.E. Correa, *Proteins*, **7,** 366 (1990).
7. J. Skolnick and A. Kolinski, *Science*, **250,** 1121 (1990).
8. E.O. Purisima and H.A. Scheraga, *Biopolymers*, **23,** 1207 (1984).
9. A. Kolinski, M. Milik, and J. Skolnick, *J. Chem. Phys.*, **94,** 3978 (1991).
10. P.J. Flory, *Statistical Mechanics of Chain Molecules*, Interscience Publishers, New York, 1969, p. 250.
11. International Mathematical and Statistical Libraries, *IMSL Math/Library*, Houston, TX, 1989.
12. J. Stoer, in *Computational Mathematical Programming*, K. Schittkowski, ed., NATO Ansi Series, 15, Springer-Verlag, Berlin, 1985, p. 165.
13. Rik Wierenga, EMBL, Heidelberg, Germany, personal communication.
14. H. Goldstein, *Classical Mechanics*, 2nd ed., Addison-Wesley, Reading, MA, 1980, p. 143.
15. Tripos Associates, Inc., *SYBYL, Molecular Modeling Software*, version 5.4, St. Louis, MO, 1991.
16. J.S. Richardson and D.C. Richardson, in *Prediction of Protein Structure and the Principles of Protein Conformation*, G.D. Fasman, ed., Plenum Press, New York, 1989, chap. 1.