# PROBIT: A statistical approach to modeling proteins from partial coordinate data using substructure libraries

**J.J. Wendoloski and F.R. Salemme**

*Du Pont Merck Pharmaceutical Company, Wilmington, DE 19880-0228, USA*

*A program (PROBIT) has been developed that allows the reconstruction of a complete set of three-dimensional protein coordinates from $\alpha$-carbon coordinates. The program generates a statistical measure of polypeptide conformational behavior for substructures in a defined structural context from a library of highly refined protein structures. These statistics provide a prescription for substructure substitution from the database to allow regeneration of the complete protein structure.*

*Keywords: homology building, protein design, rotamers, structure evaluation, computer modeling*

## INTRODUCTION

Protein engineering and rational drug design technologies rely on a detailed structural knowledge of proteins at the atomic level.[1,2] Although it would be desirable to predict protein structures directly from amino acid sequence data, this information is currently accessible only through experimental methods of X-ray crystallography and NMR. Nevertheless, many studies of protein structure have demonstrated architectural principles that recur on many levels, potentially making it possible to generate accurate models of unknown molecules from motifs present in known protein structures. These precedents motivate the development of computer methods that can make extensions from partial coordinate data that has been derived experimentally, or that can generate structures of homologous or mutated protein sequences for which experimental structural data is unavailable.

PROBIT uses a database of high-resolution crystallographic structures to rebuild other protein structures. More specifically, the program generates statistics about the local conformational properties of a substructure as reflected in a library of highly refined protein structures. The relative measures of statistical reliability then provide a global prescription for reconstructing the protein from the local fragments.

PROBIT incorporates many of the features of the program FRAGLE,[3] which was developed to fit X-ray electron density maps. FRAGLE was interfaced with FRODO,[4] and allowed the interactive retrieval and display of backbone fragments from a group of highly refined structures, together with the retrieval and display of alternative amino acid side-chain rotamers, either from a library of rotamers,[5] or as part of a backbone structure retrieved from the structural database.[6] In addition, FRAGLE incorporated a subroutine that allowed some statistical properties of substructures to be evaluated to aid interactive map fitting. PROBIT incorporates the FRAGLE structural database,[3] but is a stand-alone program that allows automatic reconstruction of protein backbone and side-chain rotamers on input of $\alpha$-carbon ($C_\alpha$) coordinates.

## BACKBONE RECONSTRUCTION

Starting from a set of $C_\alpha$ positions, it has been shown[3,7] that it is possible to search a structural database and extract local segments of backbone centered around a target residue that accurately match within a preset tolerance (default 1.0 Å). Experiments with PROBIT show that a window size of 5–9 residues is typically adequate, so that the window size is dynamically adjusted in program operation to the largest value that will produce a minimum number of matches (default 30) between the target and fragments in the database. Local segments of backbone must be merged to produce a complete backbone. Since merging separately retrieved fragments often produces poor geometry at the junctions, a smoothing operation must be provided. In PROBIT, fragments centered on each $C_\alpha$ are retrieved, then coordinates of all fragments are averaged to produce the final backbone structure. While more complex smoothing approaches have been

Address reprint requests to Dr. Wendoloski at Sterling Winthrop Pharmaceuticals Research Division, 9 Great Valley Parkway, Malvern, PA 19355, USA.
Received 28 April 1991; accepted 8 May 1991

**Table 1. RMS differences with X-ray structures**

| | Crambin | Ribonuclease | L7/L12 |
|---|---|---|---|
| All | 1.40 | 1.67 | 1.15 |
| $C_\alpha$ | 0.03 | 0.04 | 0.05 |
| Main | 0.35 | 0.38 | 0.31 |
| Side | 2.07 | 2.38 | 1.69 |

proposed,[8] the PROBIT-built backbones, followed by regularization with molecular mechanics, give low root-mean-square errors when compared to X-ray coordinates in control experiments (Table 1).

## SIDE-CHAIN RECONSTRUCTION

Ponder and Richards[5] have shown most recently that internal residues in proteins are distributed over a restricted set of well-defined rotamer conformations. Moreover, McGregor et al.[9] have shown that rotamer substrates for specific amino acids may depend upon their incorporation in a given type of secondary structure. The side-chain reconstruction strategy used in PROBIT basically incorporates a statistical generalization of these observations. Beginning from a set of backbone atom positions generated in the first pass of PROBIT, the program defines a search probe that includes a local segment of backbone centered around a target residue. The database is searched for occurrences of the central amino acid in polypeptide conformational contexts that match the target within a preset tolerance (default 1.0 Å). Typically a window size of 5–9 residues is adequate, with the window size adjusted to the largest one that will produce a minimum number of matches (default 30) in the database. Once the required number of matches is obtained, the resulting segments are analyzed to determine the rotamer populations for the target residue. These populations are taken to define a probability distribution for the residue rotamers within the local structural context in the protein, as defined by the search probe. The resulting probability distributions could, in principle, be used directly to determine which rotamer conformation is appropriate in a local context. However, for residues with multiple rotamer states of similar probability, ambiguities may exist. Clearly in such cases, the ambiguity must be resolved in ways that minimize steric interference with other residues whose rotamer states are better determined statistically in their corresponding contexts. In contrast, construction schemes that simply reconstruct the protein sequentially work poorly, since a single improper residue placement affects the packing of many succeeding residues. PROBIT provides a solution to this problem by first calculating the frequency distributions for all target residues in the sequence without substituting in any rotamers. The target residues are then substituted in order of decreasing probability, using the highest probability rotamer for each residue. Each rotamer–residue combination is checked for unfavorable van der Waals interactions, and eliminated if any are found. Residues for which the top probability rotamer is eliminated can have either the next highest probability rotamer substituted, or the residue can be resorted in the list based on the remaining residue–rotamers.

Initial control experiments comparing PROBIT reconstructions with X-ray structures indicated that aromatic residues, particularly when clustered together, were occasionally misplaced. This in part reflects the nearly equal distribution of rotamer states that frequently occurs for aromatic residues in many structural contexts. Tests of alternative strategies to pack these residues indicated that simply leaving the aromatics as the last class of residue–rotamers to be substituted was an effective strategy. This result suggests that aromatic residue–rotamers are often determined by packing constraints of the nonaromatic residues. PROBIT optionally allows either normal or special aromatic residue building, contingent upon the examination of the rotamer distributions from aromatic residues in the specific protein under reconstruction.

The higher probability residue–rotamer combinations, which are rebuilt first in PROBIT reconstruction, appear to function as "keystone" residues that act to conformationally restrict the rotamer states of less deterministic residues. At the same time, the resulting statistical behavior provides a measure of reliability in the placement of individual reconstructed side chains.

**Table 2. Percentage $\chi$-angles correct versus X-ray***

| | Crambin | Ribonuclease | L7/L12 |
|---|---|---|---|
| $\chi^1$ | 59 | 53 | 70 |
| $\chi^{1+2}$ | 54 | 39 | 59 |

*Correct within 40°

## APPLICATIONS AND RESULTS

PROBIT has been used successfully for automatic generation of coordinate sets for protein refinement, based on initial $C_\alpha$ chain traces, as well as in numerous applications in homology building and protein engineering. Visual comparison of reconstructed control molecules with X-ray coordinates generally show good correspondence for internal residues, although as expected, solvent-exposed residues like lysine are predicted poorly. Results from three control experiments are given in Tables 1 and 2. Overall, PROBIT provides agreement with X-ray results similar to that obtained by several alternative methods described elsewhere. These include methods that use structural templates,[10] as well as more theoretical automated approaches using empirical force fields combined with simulated annealing,[11] combinations of structural data and empirical forcefields,[12–14] or Monte Carlo sampling with simulated annealing.[8] However, an important aspect of the PROBIT approach is that it provides a statistical estimate of residue placement accuracy that may be incorporated usefully in applications like structure refinement against X-ray or NMR-distance constraint data.

The program is a licensed product that can be obtained by contacting the authors. Since the present implementation utilizes a number of FRODO subroutines,[4,15] the program can be distributed only to holders of valid FRODO licenses. A version that does not have this restriction is anticipated.

## REFERENCES

1 Hol, W.G.J. Applying knowledge of protein structure and function. *Tibtech* 1987, **5**, 137–143

2 Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E., and Thornton, J.M. Knowledge-based prediction

of protein structures and the design of novel molecules. *Nature* 1987, **326**, No. 6111, 347–352

3 Finzel, B.C., Kimatian, S., Ohlendorf, D.H., Wendoloski, J.J., Levitt, M., and Salemme, F.R. Molecular Modeling with Substructure Libraries Derived from Known Protein Structures. In *Crystallographic and Modeling Methods in Molecular Design* (S. Ealick and C. Bugg, eds.) Springer Verlag, New York, 1990, 175–189

4 Jones, T.A. Interactive computer graphics:FRODO. *Meth. Enzymol.* 1985, **115**, 157–171

5 Ponder, J.W. and Richards, F.M. Tertiary templates for proteins. *J. Mol. Biol.* 1987 **195**, 775–791

6 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 1977, **112**, 535–542

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

8 Holm, L. and Sander, C. Database algorithm for generating protein backbone and side chain coordinates from a $C_\alpha$ trace. Application to model building and detection of coordinate errors. *J. Mol. Biol.* in press

9 McGregor, M.J., Islam, S.A., Sternberg, M.J.E. Analysis of the Relationship Between Side-Chain Conformation and Secondary Structure in Globular Proteins. *J. Mol. Biol.* 1987 **198**, 295–310

10 Reid, L.S., and Thornton, J.M. Rebuilding flavodoxin from $C_\alpha$ Coordinates: A test study. *Proteins: Structure, Function, and Genetics* 1989, **5**, 170–182

11 Correa, P.E. The Building of protein structures from $\alpha$-Carbon coordinates. *Proteins* 1990, **7**, 366–377

12 Summers, N.L., Carlson, W.D., and Karplus, M. Analysis of Side-Chain Orientations in Construction of Side Chains in Homologous Proteins. *J. Miol. Biol.* 1987, **196**, 175–198

13 Summers, N.L. Monsanto Researchers Apply CAMD Methods in Globular Protein Research. *Chemical Design Automation News* 1989, **4**(12) 1, 12–16

14 Summers, N.L. and Karplus, M. Construction of Side Chains in Homology Modeling: Application to the C-terminal lobe of Rhizopuspepsin. *J. Mol. Biol.* 1989, **210**, 785–811

15 Pflugrath, J.W., Saper, M.A., and Quiocho, F.A. In: *Methods and Applications in Crystallographic Computing* (S. Hall and T. Ashida, eds.) Oxford University Press, London, 1984, 404