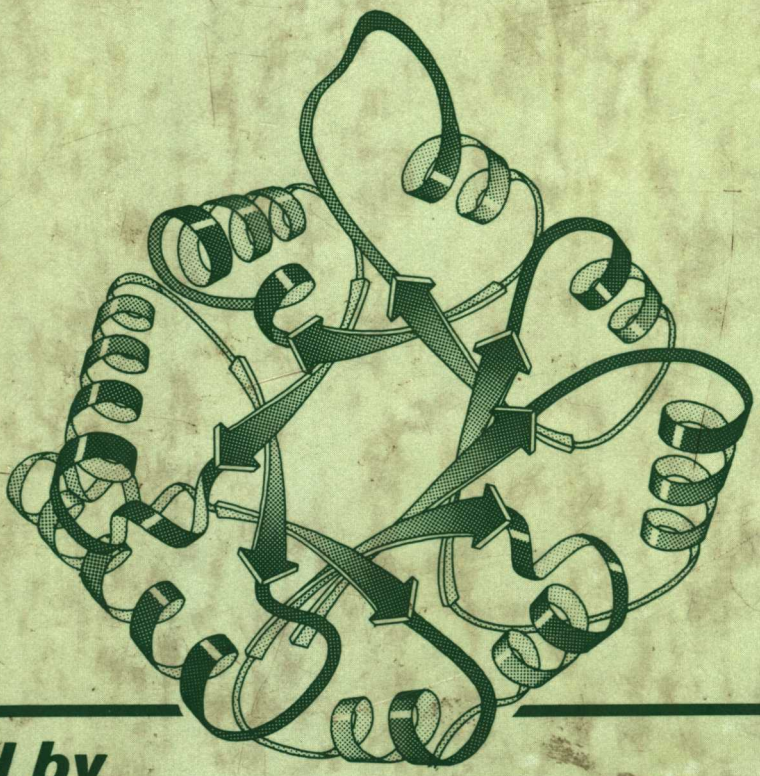
PREDICTION OF PROTEIN STRUCTURE AND THE PRINCIPLES OF PROTEIN CONFORMATION



Edited by
Gerald D. Fasman

Prediction of Protein Structure and the Principles of Protein Conformation

Edited by
Gerald D. Fasman
Brandeis University
Waltham, Massachusetts

Plenum Press • New York and London

Library of Congress Cataloging in Publication Data

Prediction of protein structure and the principles of protein conformation / edited by Gerald D. Fasman.

p. cm.

Includes bibliographies and index.

ISBN 0-306-43131-9

1. Proteins—Structure—Mathematical models. 2. Proteins—Conformation. I. Fasman, Gerald D.

QP551.P655 1989 89-8555

574.19'296'011—dc20 CIP

Cover illustration: Ribbon drawing of triose phosphate isomerase (end view) as the classic example of a singly wound parallel β barrel. Figure 44 in Chapter 1, courtesy of Drs. Jane S. Richardon and David C. Richardson.

A Division of Plenum Publishing Corporation

233 Spring Street, New York, N.Y. 10013

Printed in the United States of America

Contributors

Tom Alber • Institute of Molecular Biology, University of Oregon, Eugene, Oregon 97403. Present Address: Department of Biochemistry, University of Utah School of Medicine, Salt Lake City, Utah 84132

J. Fernando Bazan • Department of Biochemistry and Biophysics, University of California, San Francisco, California 94143

Peter Y. Chou • 3893 Ross Road, Palo Alto, California 94303

Fred E. Cohen • Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, California 94143–04460

‡A. J. Cross • Biosym Technologies, Inc., San Diego, California 92121

Johann Deisenhofer • Department of Biochemistry, and Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas 75235

Gilbert Deléage • Laboratory of Biological Chemistry, Claude Bernard University of Lyon I, 69622 Villeurbanne Cedex, France

Russell F. Doolittle • Center for Molecular Genetics, University of California, San Diego, La Jolla, California 92093

Jonathan E. Dworkin • Department of Biological Chemistry, Milton S. Hershey Medical Center, Pennsylvania State University, Hershey, Pennsylvania 17033

David Eisenberg • Molecular Biology Institute, and Departments of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, California 90024

Gerald D. Fasman • Graduate Department of Biochemistry, Brandeis University, Waltham, Massachusetts 02254

Janet Finer-Moore • Department of Biochemistry and Biophysics, University of California, San Francisco, California 94143

J. Garnier • Laboratory of Physical Biochemistry, University of Paris-Sud, 91405 Orsay, France. Present Address: Protein Engineering Unit, INRA, 78350 Jouy-en-Josas, France

‡Deceased.

A. T. Hagler • Biosym Technologies, Inc., San Diego, California 92121; and Agouron Institute, La Jolla, California 92137.

Robert Huber • Division for Structural Research, Max Planck Institute for Biochemistry, D-8033 Martinsried, West Germany

Fritz Jähnig • Max Planck Institute for Biology, D-7400 Tübingen, West Germany

‡Emil T. Kaiser • Laboratory of Bioorganic Chemistry and Biochemistry, Rockefeller University, New York, New York 10021

Irwin D. Kuntz • Department of Pharmaceutical Chemistry, School of Pharmacy, University of California, San Francisco, California 94143–04460

Terry P. Lybrand • Department of Chemistry, University of Houston, Houston, Texas 77004. Present Address: Department of Medicinal Chemistry, University of Minnesota, Minneapolis, Minnesota 55455

D. H. J. Mackay • Biosym Technologies, Inc., San Diego, California 92121

J. Andrew McCammon • Department of Chemistry, University of Houston, Houston, Texas 77004

Harmut Michel • Division for Molecular Membrane Biochemistry, Max Planck Institute for Biophysics, D-6000 Frankfurt 71, West Germany

Kozo Nagano • Faculty of Pharmaceutical Sciences, University of Tokyo, Tokyo 113, Japan

Peter Prevelige, Jr. • Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

David C. Richardson • Department of Biochemistry, Duke University, Durham, North Carolina 27710

Jane S. Richardson • Department of Biochemistry, Duke University, Durham, North Carolina 27710

B. Robson • Proteus Biotechnology Ltd., Marple, Cheshire SK6 6AB, England; and Epistron Peptide and Protein Engineering Research Unit, University of Manchester, Manchester M13 9PT, England

Neil K. Rogers • Laboratory of Molecular Biophysics, Department of Zoology, University of Oxford, Oxford OX1 3QU, United Kingdom

George D. Rose • Department of Biological Chemistry, Milton S. Hershey Medical Center, Pennsylvania State University, Hershey, Pennsylvania 17033

Bernard Roux • Laboratory of Biological Chemistry, Claude Bernard University of Lyon I, 69622 Villeurbanne Cedex, France

John Rubin • Genentech, Inc., South San Francisco, California 94080

Robert M. Stroud • Department of Biochemistry and Biophysics, University of California, San Francisco, California 94143

Gerald Stubbs • Department of Molecular Biology, Vanderbilt University, Nashville, Tennessee 37235

‡Deceased.

Morgan Wesson • Molecular Biology Institute, and Departments of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, California 90024

William Wilcox • Molecular Biology Institute, and Departments of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, California 90024

Chung F. Wong • Department of Chemistry, University of Houston, Houston, Texas 77004

Preface

The prediction of the conformation of proteins has developed from an intellectual exercise into a serious practical endeavor that has great promise to yield new stable enzymes, products of pharmacological significance, and catalysts of great potential. With the application of prediction gaining momentum in various fields, such as enzymology and immunology, it was deemed time that a volume be published to make available a thorough evaluation of present methods, for researchers in this field to expound fully the virtues of various algorithms, to open the field to a wider audience, and to offer the scientific public an opportunity to examine carefully its successes and failures. In this manner the practitioners of the art could better evaluate the tools and the output so that their expectations and applications could be more realistic.

The editor has assembled chapters by many of the main contributors to this area and simultaneously placed their programs at three national resources so that they are readily available to those who wish to apply them to their personal interests. These algorithms, written by their originators, when utilized on PCs or larger computers, can instantaneously take a primary amino acid sequence and produce a two- or three-dimensional artistic image that gives satisfaction to one's esthetic sensibilities and food for thought concerning the structure and function of proteins. It is in this spirit that this volume was envisaged.

Thanks are due to Pamela Gailey for her assistance in the handling of the manuscripts and to Mary Born of Plenum Press, whose editorial help was always forthcoming. I owe a great debt to the staff of the Gerstenzang Science Library at Brandeis University for being so generous with their time in assisting me. Of course, the main credit for this volume belongs to the authors of the various chapters, who have labored to make their work available to the public.

The time is approaching when a new protein will be designed on the drawing board, using predictive algorithms, and its subsequent synthesis, via cloning or peptide coupling, will offer new and interesting challenges for biochemists and molecular biologists.

Gerald D. Fasman

Waltham, Massachusetts

Contents

1

Principles and Patterns of Protein Conformation

Jane S. Richardson and David C. Richardson

Jane S. Richardson and David C. Richardson • Department of Biochemistry, Duke University, Durham, North Carolina 27710.

I. INTRODUCTION

The raw materials of protein structure are the detailed geometry and chemistry of the polypeptide and side chains plus the solvent environment. The end result is a complex tapestry of details organized into a biologically meaningful whole: a variation on one of a few harmonious themes of three-dimensional structure. For the purposes of prediction we are not concerned primarily with either of the endpoints of this process but with the logical connection between the two. Therefore, we summarize what is known of that logical connection into a set of guiding principles: hydrophobicity, hydrogen bonding, handedness, history, and the tension between hierarchy and interrelatedness. In addition, we consider particularly relevant features of the starting and ending states. However, one should bear in mind, as cartooned in Fig. 1, that our abilities to follow a protein through this remarkable transition are still rather limited in both the experimental and the theoretical realms.

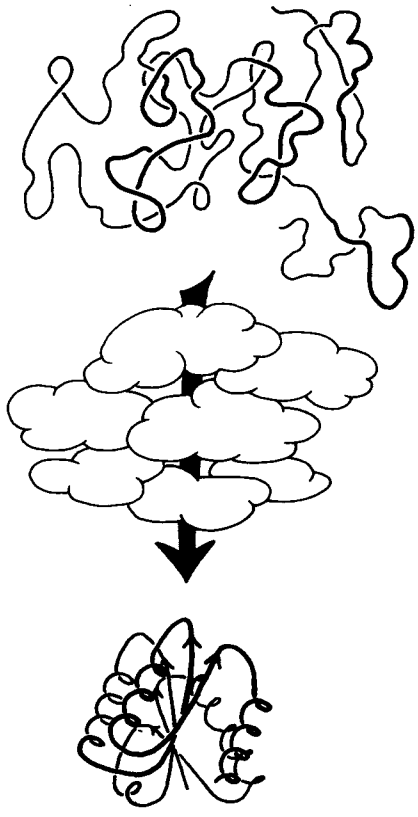


Figure 1. Protein folding.

Fortunately, there is a large and detailed body of evidence about the final folded state, from x-ray crystallography and now increasingly from two-dimensional NMR, most of which is available from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977). Appendix 1 lists the proteins used in assembling the data and figures for this chapter with their Brookhaven identification codes.

II. THE CONSTRAINTS AND OPPORTUNITIES INHERENT IN THE POLYPEPTIDE CHAIN

Because of the remarkable self-assembly capabilities shown by many proteins for refolding *in vitro,* we know that most aspects of protein configuration derive, in the final analysis, from the properties of the particular sequence of amino acids that make up its polypeptide chain. These properties include the characteristics of both the individual side chains and the polypeptide backbone, which exerts its influence in ubiquitous and sometimes rather subtle ways.

The geometric parameters of the peptide unit and the α carbon are illustrated in Fig. 2 (Momany *et al.*, 1975), and the two are strung together into the familiar polypeptide backbone in Fig. 3. The bond lengths and angles shown are the local minimum-energy values around which the structure fluctuates, both as a function of time for a given bond or angle and also statistically among the total set of such bonds or angles. For bond lengths the restoring forces are large, the time scale is very short, and the range of variation is quite small. Bond angles are somewhat looser, and dihedral or torsion angles are looser still.

Aside from its linear connectivity and steric volume, the most pervasive and significant influences of the polypeptide chain itself on protein conformation are the hydrogen-bonding capabilities of the peptide (treated in Section V) and the handedness imposed by the asymmetry of the α carbon. The convention shown in Fig. 4 (the "CORN crib") allows one to recognize the correct L-amino-acid handedness when dealing with physical models, stereo figures, or molecular-graphics displays: if one looks down on the α carbon from the direction of the hydrogen, the other substituents should read "CO-R-N" in clockwise order, where R stands for the R group of the side chain, CO for the peptide carbonyl, and N for the peptide nitrogen. Later we will investigate the unique role glycine plays because of its lack of handedness (Section V.A) and the influence L-amino acids exert on large-scale handedness phenomena at all levels of protein structure (Section VII).

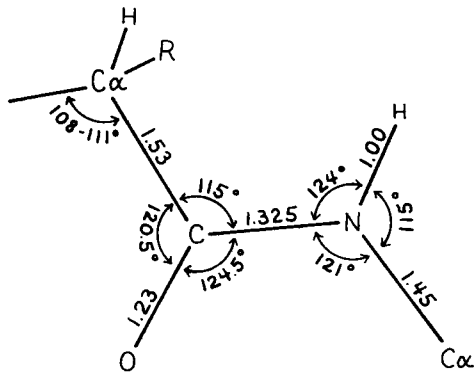


Figure 2. Numerical values for the bond lengths and angles of a peptide and an α carbon.

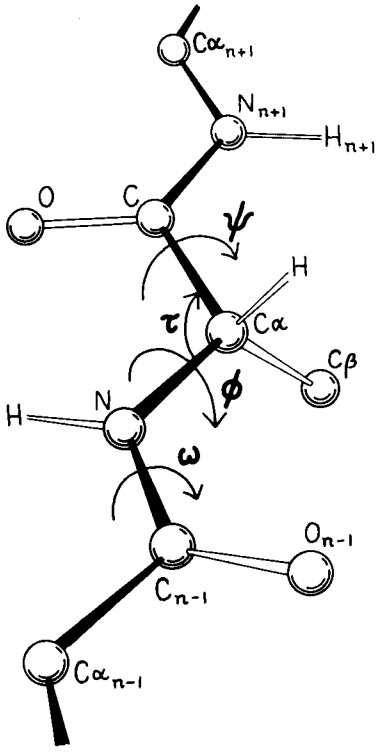


Figure 3. A key to nomenclature for the atoms of the polypeptide chain, the tetrahedral bond angle τ, and the backbone dihedral angles ϕ, ψ, and ω.

The other convention needed for understanding conformational details is the rule for assigning direction and numerical values to dihedral angles, as shown in Fig. 5. A dihedral angle involves four successive atoms—A, B, C, and D—and the three bonds joining them. If you look directly down the length of the central bond joining atoms B and C (fortunately, the answer is the same as viewed from either end of this bond) and put the atom nearest you (A) at 12 on the clock face, then the clock position of the far atom (D) reads out the angle. By convention, dihedral angles are assigned in the range $-180°$ to $+180°$ with the clockwise direction being positive. Thus, the dihedral angle shown in Fig. 5 is about $+35°$.

Assuming ideality for the rest of the geometry, then the three backbone dihedral angles per residue (ϕ, ψ, and ω) plus the dihedral angles χ_1, χ_2, . . . out the side chain provide a complete description of the local conformation. This description is ideal for comparing short pieces of structure, since it is independent of reference frame, but it is not workable for specifying global conformation because even very small round-off errors accumulate drastically. In practice, just ϕ and ψ suffice for the main chain, because the partial-double-bond

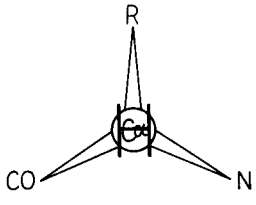


Figure 4. The "CORN crib": a mnemonic for determining the handedness of an amino acid. Looking at the α carbon from the direction of the hydrogen, the other substituents should read CO (carbonyl), R (side chain), and N (backbone NH) in clockwise order for a biologically appropriate L-amino acid.

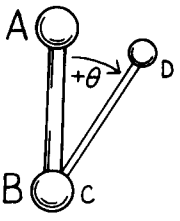


Figure 5. Illustration of the standard convention (IUPAC, 1970) for measuring dihedral angles. The dihedral, or torsion, angle around a bond B–C is defined by the relative positions of the four atoms A, B, C, and D. Looking down the B–C bond, atom A is placed at 12 o'clock, and atom D measures the dihedral angle: plus if clockwise, as in this example (about +35°), and minus if counterclockwise.

character of the peptide keeps ω very close to flat, with the two successive α carbons and the C, O, N, and H between them all lying in one plane. ω is almost always within about 10° of 180°, which is the fully extended or "*trans*" conformation. The curled-up "*cis*" conformation of ω at or near 0° is observed about 10% of the time for proline (see Section V.B for details) and extremely rarely for any other amino acid.

Since φ and ψ form a virtually complete description of backbone conformation, a two-dimensional plot of them (known as a Ramachandran plot) is an important type of representation. We use such plots to illustrate properties of repeating conformations, single residues, or two successive residues.

Regions of φ,ψ space are generally named after the conformation that results if they are repeated. In Fig. 6, the major regions are the right-handed α-helical cluster in the lower left, near −60°,−40°; the broad region of extended β strands in the upper left quadrant (centered around −120°,140°); and the sparsely populated left-handed α-helical region in the upper right near +60°,+40°. Vacant areas are conformations that place atoms unfavorably close together within the dipeptide unit: near 0°,0° the oxygen of residue $n - 1$ bumps the carbonyl C of residue n. The asymmetry of the plot results from collisions with C_β. Within each conformational region there can be significant differences, such as parallel versus antiparallel β, widely varying degrees of β-sheet twist, and extended collagen-type helix all lying within the "β" region. The "bridge" area across $\psi = 0°$ between the α and β regions should be unfavorable based on a hard-sphere model, and yet it is in fact rather well populated; since the bump involved is between successive amide groups, it could be relieved either if the bond angle τ at the α carbon can stretch wider than tetrahedral or if the amide hydrogen is a bit "soft."

A more sophisticated energy calculation for the dipeptide around a central C_α (usually Ala) can do a fairly good job of matching the observed φ,ψ distribution (Anderson and Hermans, 1988). In a way this is rather surprising, because such a calculation leaves out both the favorable and the unfavorable effects of long-range interactions of the backbone as well as specific side-chain effects. One of the more remarkable properties of the repetitive secondary structures observed in proteins is that the optimum φ,ψ values and the permissible range for good long-range H-bonding and steric fit are so close to the optimum and range for favorable dipeptide conformations. Figure 7 shows a φ,ψ plot for residues in nonrepetitive loops, and one sees that the match to Fig. 6 is rather close in spite of the absence of helix and sheet. The presumption is that this neat match is what has, for instance, so strongly selected for the occurrence of right-handed α helices rather than for any of the slightly different versions such as 3_{10}, π, or left-handed α helices.

One should recognize that this fit of local and long-range preferences assumes that the influence of side-chain interactions is so variable as to cancel out on the average. This would not necessarily be true for highly repetitive amino-acid sequences, and there is every reason to suppose that these produce a different range of conformations. This has been convincingly demonstrated for elastin (poly VPGVG: Cook *et al.*, 1980) and collagen (Rich and Crick, 1961), for instance, and has been proposed for a number of other cases. Rather surprisingly, in

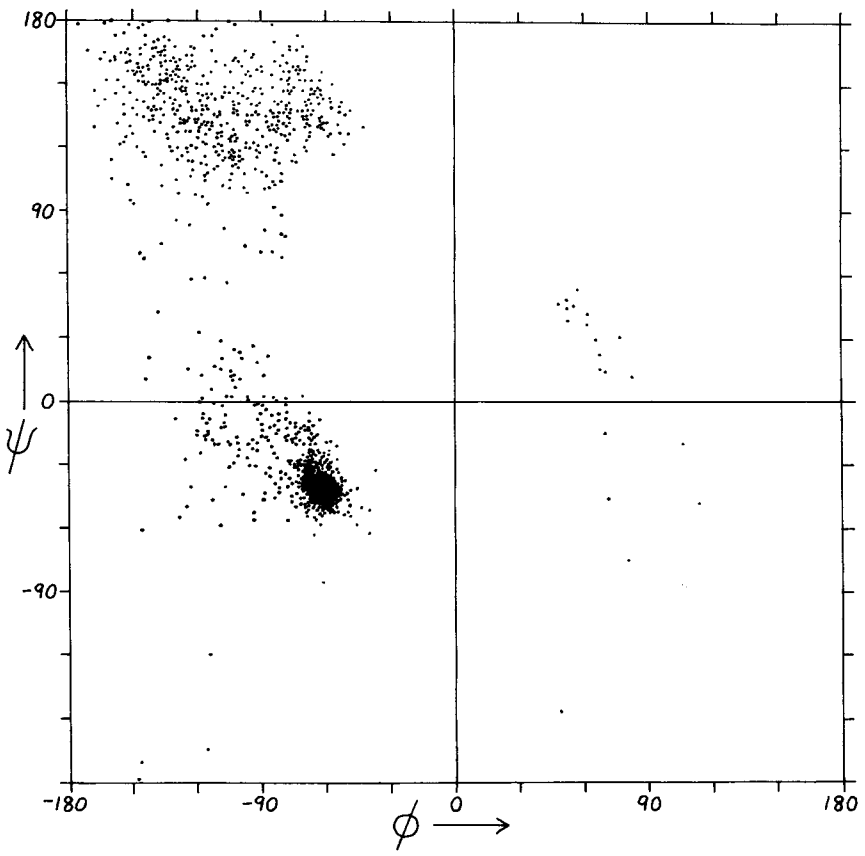


Figure 6. Plot of ϕ,ψ values found for all residues (excluding Gly and Pro) in a representative sample of highly refined x-ray structures at resolutions of 1.2 to 1.8 Å: 4DFR, 1ECA, 4FXN, 1INS, 2MHR, 1OVO, 1PCY, 5PTI, 2RHE, 5RXN, 2SGA, 1SN3.

the known globular protein structures there are no recognized examples of structural features that involve a cyclic repeat of two, three, or four ϕ,ψ values, except for the case of small alternating ϕ,ψ perturbations producing bends in β ribbons (see Section III.B).

If the broad conformational regions (such as "β" or "extended" taken to include most of the upper left quadrant of the ϕ,ψ plot) are considered in more detail, then some systematic differences emerge between repetitive and nonrepetitive conformations. Figure 7 shows a strong clustering around the polyproline conformation (near $\phi = -60°$, $\psi = 140°$) in spite of the fact that both Pro and Gly were omitted from the plot; there are also somewhat more points in the bridge region near $\phi = -90°$, $\psi = 0°$. Why are these conformations so commonly observed when they are neither at the local minimum for single amino acids nor very favorable for repeating structures? The answer appears to be that poly-Pro and 3_{10} conformations place successive carbonyls approximately perpendicular to each other rather than parallel as in α helix or antiparallel as in β strands. Figure 8 shows the angle between CO_i and CO_{i+1} as a function of ϕ and ψ. (Note: this function can also be expressed as the angle between successive peptide NH groups, which is a more useful construct for NMR.) Many of the common transitions between pieces of secondary structure (such as α–β connections) involve a 90°

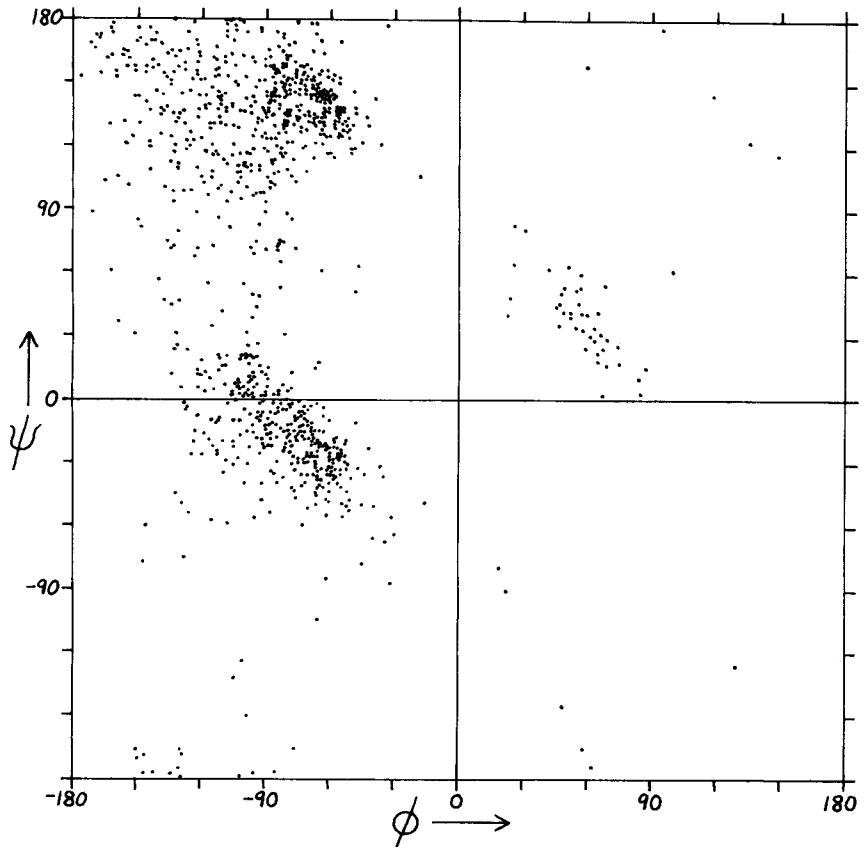


Figure 7. Plot of ϕ,ψ values for about 1000 residues in nonrepetitive structure (again excluding Gly and Pro) in protein crystal structures at about 2 Å resolution or better. Compared with Fig. 6, the same regions are populated, but there is relatively more emphasis of the polyproline, 3_{10}, $L\alpha$, and "bridge" conformations and less of α and β.

change in carbonyl direction, so if they are accomplished in a small number of residues these will usually include at least one of these "perpendicular" residues. This same distinction holds, of course, for the "left-handed" conformations with positive ϕ angles (which, unfortunately for terminological clarity, are on the right side of the standard ϕ,ψ plot). More than half of the "$L\alpha$" glycines are actually $L3_{10}$, so that their surrounding carbonyls are nearly perpendicular. Most helices and β strands begin and end with a residue in one of the perpendicular conformations (e.g., Fig. 9), and a tight turn requires two of them. The various perpendicular conformations act as the punctuation between secondary structures or as the creases that fold a polypeptide into the elaborate origami construction of a globular protein.

A more subtle influence of backbone geometry arises from the nonorthogonality of its rotations. If the bond angles were 90° and the bond lengths such that two bonds separated in the sequence could be collinear with each other, then it would be possible for some external subsections of the chain to move completely independently of the rest of the protein. As it is, such conditions can only be satisfied very approximately. The simplest example is coupled rotation of the angles at both ends of a single peptide, which leaves the net chain direction