PROTEINS

Structures and Molecular Principles

THOMAS E. CREIGHTON
Medical Research Council
Laboratory of Molecular Biology
Cambridge, England

W. H. FREEMAN AND COMPANY NEW YORK

Cover and text design: Edward A. Butler

Library of Congress Cataloging in Publication Data
Creighton, Thomas E., 1940 -
Proteins.

Includes bibliographies and index.
1. Proteins. I. Title. [DNLM: 1. Proteins. QU 55 C914p]

QP551.C737 1984 574. 19'245 83-19541

ISBN 0-7167-1566-X

Printed in the United States of America

1 2 3 4 5 6 7 8 9 0

PREFACE

While working with proteins during the past 20 years, I often felt the need for a comprehensive text on their physical and chemical properties. Such a text was simply not available. A search of the literature for information about any of the many diverse aspects of proteins required countless hours. No less daunting was the task of reconciling the different, often conflicting, explanations given for these properties. This book attempts to provide such information within a coherent framework.

Proteins have been studied for so long by so many different people that there is no shortage of material in the literature. However, because it is scattered very widely, from journals of physical chemistry to those of microbiology and genetics, the information needed to be gathered together and then sifted, sorted, and rationalized. This process revealed that most of the information is incomplete; in many cases, research of a particular problem apparently ended when only the vague outline of the answer became evident. Perhaps those working directly on such a problem see the answer clearly but usually do not bother to write a definitive account, presumably moving on to newer, more exciting fields. Much of the material is also contradictory. When a choice between conflicting views became necessary, I have attempted to write as definitive an account as possible and have presented the view that is both more plausible and more compatible with all other knowledge about proteins.

Upon some aspects, such as protein stability, this volume presents a very personal point of view, which does not represent a consensus. I have been driven to this only when there was no plausible treatment of the problem in the literature. Otherwise, I have attempted to convey the conventional wisdom that has evolved, while trying to analyze its flaws, point

out its deficiencies, and present it in the most convincing manner. Although this volume is as comprehensive as possible within the limits of the space and information available, it cannot claim to be complete, for there are still many aspects of protein structure and function that are not understood.

The reader is expected to have primarily a background in biology, with some knowledge of biochemistry, genetics, and cell biology. Accordingly, this book addresses at the lowest level the physical and chemical properties of proteins that are least likely to be intuitively obvious to such a reader; the treatment is not rigorous but is often empirical, presented in a manner meant to convey principles without obscuring them in mathematical derivations. (The reader is directed to textbooks of biophysical chemistry for such rigorous derivations.) To a chemist or physicist with no biology background, many of the discussions of physical chemistry may seem trivial, but the significance of many of the biological properties should not be overlooked. (A complete explanation of all the biological phenomena would have required a volume many times as large; and there are very many biologic textbooks that present such information clearly and comprehensively.)

The text is organized on a combined basis of the chronological events in the life of a protein and of structural and functional hierarchy. Biosynthesis is discussed early (Chapter 2), although first it is necessary to introduce the parts that make up proteins (Chapter 1); the volume ends with a discussion of degradation (Chapter 10). The first three chapters deal with properties of linear polypeptide chains, where only the primary structure is immediately relevant, as in biosynthesis, and includes a discussion of the origins of contemporary primary structures (Chapter 3). The next three chapters describe the three-dimensional aspects of proteins, starting with a brief description of the forces controlling this level of structure (Chapter 4), continuing with their effects on conformational properties of the polypeptide chain (Chapter 5), and ending with a description of the folded structures of globular proteins (Chapter 6). The subsequent three chapters describe the consequences of the folded structures of proteins, including their dynamic properties in solution (Chapter 7), their interactions with other molecules (Chapter 8), and their ability to catalyze chemical reactions and interconvert different forms of energy (Chapter 9).

Documentation of each statement in the text with a reference is not necessary for a volume at this level. Instead, a few of the most pertinent references are listed at the end of each section, for convenience and effeciency; the more general references are listed after the introduction to a chapter or to a group of sections.

I have endeavoured to select references on the basis of their usefulness to the reader, rather than the gratification of scientists' need for recognition. Consequently, the choice of references should not be taken as evidence of their contribution to advancing knowledge but simply reflects that they are useful compendiums of data, present coherent discussions,

or simply have nice illustrations. Only very exceptionally important references more than ten years old are cited; some of the most recent references were included simply to provide a recent entry into the literature of a particular field, as a starting point for tracing back earlier references, and as an indication of the current status of the field.

This seems an appropriate opportunity to thank those who contributed both directly and indirectly to this book. First mention must go to my colleagues and the many visitors at the Medical Research Council Laboratory of Molecular Biology, who provide an unsurpassed scientific environment, especially those with whom I have spent many hours in the canteen discussing proteins and other topics. In particular, David Goldenberg provided numerous perceptive comments and questions in such conversations; he also contributed directly to this volume by reading the entire manuscript and pointing out errors, inconsistencies and shortcomings. More remote, but no less important, contributions were made by those who provided the very important support and encouragement at the start of my career: Robert L. Sinsheimer and George Guthrie, who introduced me to molecular biology and to research while I was an undergraduate at the California Institute of Technology; and Charles Yanofsky, who patiently guided me to the Ph.D. degree at Stanford, and who provided an exemplary model of how being a great scientist need not prevent one from also being a kind and generous human being.

CONTENTS

1

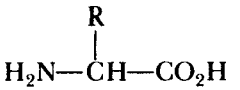
CHEMICAL NATURE OF POLYPEPTIDES

Confronting the full three-dimensional model of a typical protein, with its multitude of H, C, O, N, and S atoms, can be an awesome experience. To the uninitiated eye, it is simply an incomprehensible myriad of atoms and bonds. Nevertheless, it can be comprehended, even admired, once a few principles of protein structure are grasped. Then, many of the biological properties of the protein can be rationalized. The first step is to divide the protein up into its various components and to become acquainted with their individual properties.

Proteins fall into the general class of polymers, which are simply linear molecules built up from simple repeating units, the monomers. In the case of proteins, the monomers are the amino acids; 20 different amino acids are used.

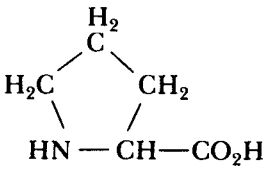
In one sense, proteins are more complex than most polymers in that 20 different monomers are used in their construction, whereas many polymers have only a single type of monomer. In another sense, however, proteins are structurally less complex: Most chemical polymers are created by polymerizing a mixture of the monomers, thereby producing a distribution of chain lengths and an approximately random sequence if more than one type of monomer is present. In contrast, proteins have precise lengths of polypeptide chain and exact sequences of the amino acids. They also are linear and unbranched. Most important, they have the property of acquiring very specific folded three-dimensional conformations, which will be described in later chapters.

Of the 20 amino acids normally used to build proteins, 19 have the general structure



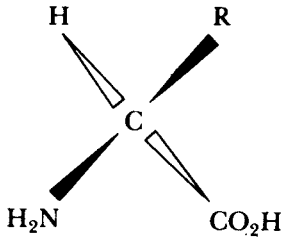
(1–1)

and differ only in the chemical structure of the side chain, R. Proline, the 20th natural amino acid, is similar but has the side chain bonded to the nitrogen atom, to give the imino acid:



(1–2)

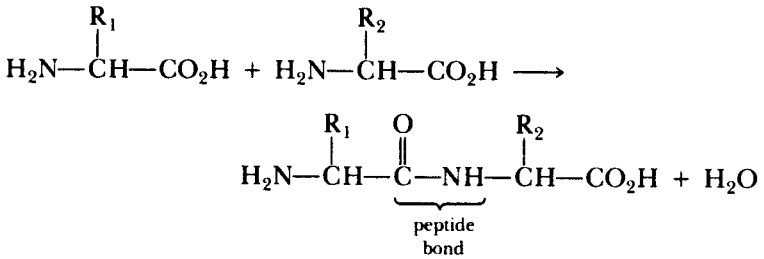
Except in glycine, where the side chain is simply a hydrogen atom, the α-carbon atom is asymmetric and always of the L isomer:



(1–3)

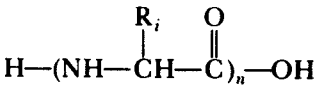
The structures of the side chains of the 20 amino acids are illustrated in Figure 1–1. Also given are their three- and one-letter abbreviations. Because the former are obvious and used widely, they will be employed here.

The 20 amino acids are assembled into proteins by linking them together via peptide bonds, as illustrated here for the condensation of two amino acids:



(1–4)

Many such amino acids, generally from 50 to 1000, are linked together in this way to form a linear polypeptide chain. The polypeptide backbone is simply a repetition of the basic amino acid unit, which is described as an amino acid **residue** when incorporated into a polypeptide chain:



(1–5)

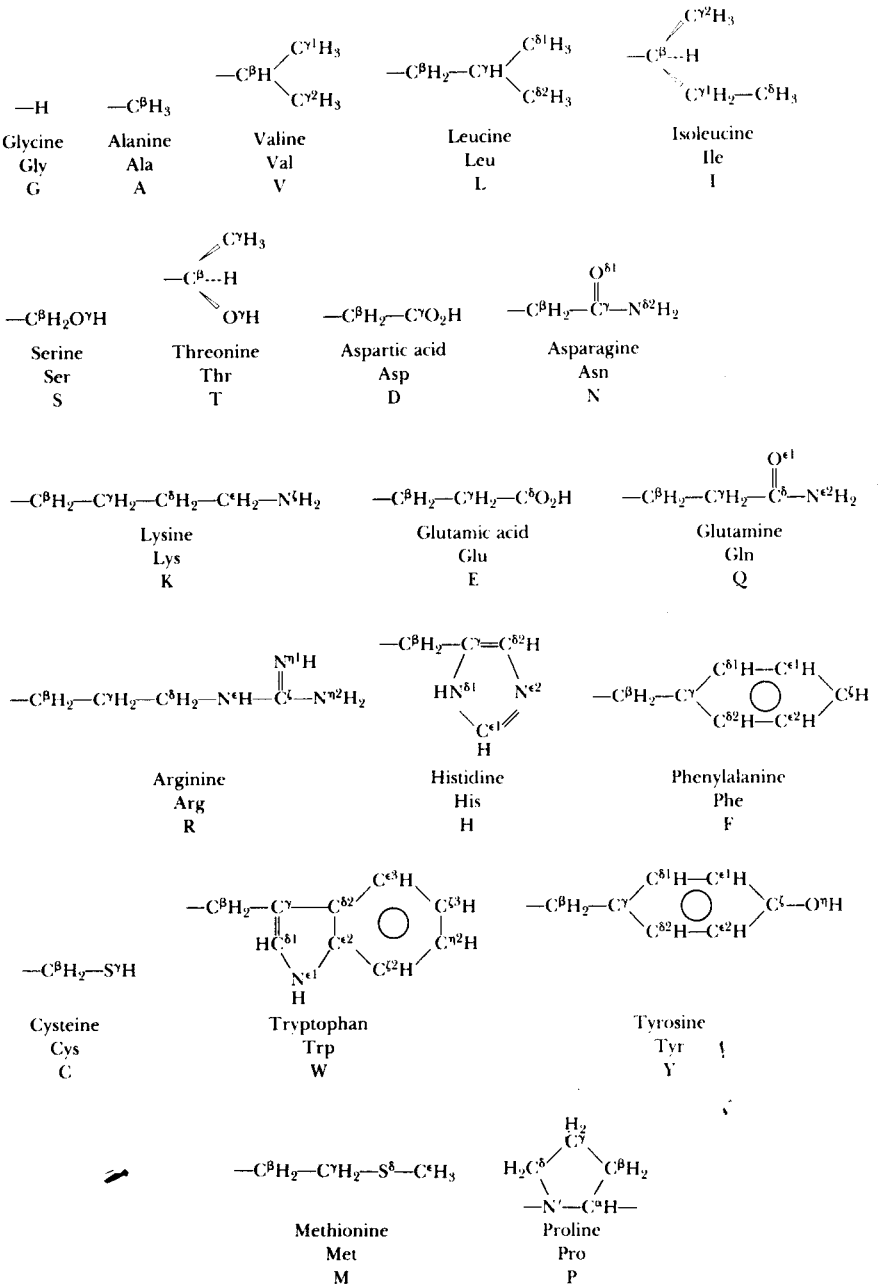


Figure 1–1
Side chains of the 20 amino acids used to synthesize proteins. The full names and the three- and one-letter abbreviations are given below the structures. The C and N atoms of the backbone are also included in the unique case of Pro. The designations of the atoms are those recommended by the IUPAC–IUB Commission on Biochemical Nomenclature. (In J. Mol. Biol. 52:1–17, 1970.)

All proteins and polypeptides have this simple basic structure and differ only in the number of amino acid residues linked together in the chain (n in Eq. 1–5) and in the sequence in which the various amino acids occur in the polypeptide chain.

It may be useful at this stage to define the various terms that are used. A **peptide** generally refers to only a small number of amino acid residues linked together, usually with a defined sequence. No particular maximum number of residues may be specified, but the term peptide is appropriate if the physical properties are generally those expected from the total of the constituent amino acids. A **polypeptide** generally refers to longer chains, but with either the sequence or the length not defined. Such polymers are often prepared by chemical polymerization of one or a few amino acids into random sequences of varying lengths. They usually have no defined conformation, or they acquire simple repetitive structures such as helices or sheets (see Chapter 4). Proteins of specific sequence are often referred to as polypeptides if they are not in a defined conformation.

The term **protein** is usually reserved for those chains with a specific sequence, length, and folded conformation. It is this class that is the subject of this volume, although useful reference will also be made to peptides and polypeptides.

Proteins, Amino Acids, and Peptides. E. J. Cohn and J. T. Edsall. Princeton, N.J., Van Nostrand-Reinhold, 1943.

X-ray studies of amino acids and peptides. R. B. Corey. Adv. Protein Chem. 4:385–406, 1948.

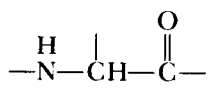
Crystal structure studies of amino acids and peptides. R. E. Marsh and J. Donohue. Adv. Protein Chem. 22:235–256, 1967.

The Structure and Action of Proteins. R. E. Dickerson and I. Geis. New York, Harper & Row, 1969.

Nomenclature of α-amino acids. IUPAC Commission on the Nomenclature of Organic Chemistry and IUPAC–IUB Commission on Biochemical Nomenclature. Eur. J. Biochem. 53:1–14, 1975.

THE POLYPEPTIDE BACKBONE

The peptide backbone consists of a repeated sequence of three atoms: the amide N, the alpha C, and the carbonyl C:



(1–6)

which are generally represented as N_i, C_i^{α}, and C_i', respectively, where i is the number of the residue, starting from the amino end.

The dimensions of the peptide group derived from three-dimensional crystal-structure analyses of small peptides are given in Figure 1–2. The maximum distance between repeating atoms in the polypeptide backbone is 3.80 Å, when the peptide bond is *trans*. In a fully extended chain consisting of many residues, the repeating units are staggered, so that the maximum linear dimension of a polypeptide with n residues is $n \times 3.63$ Å.

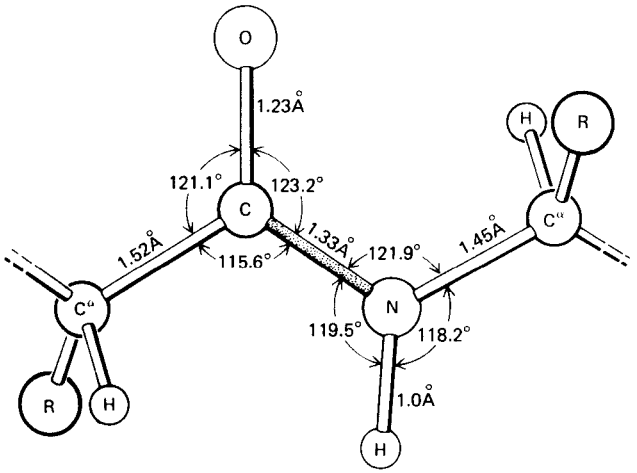
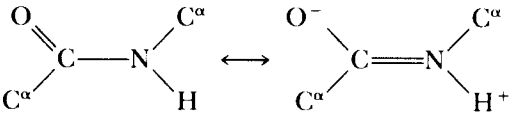


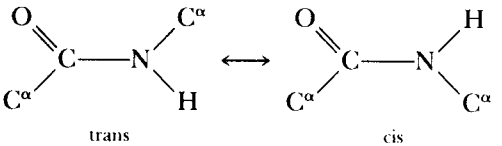
Figure 1–2
The geometry of the peptide backbone, with a trans peptide bond, showing all the atoms between two C^{α} atoms of adjacent residues. The peptide bond is stippled. The dimensions given are the averages observed crystallographically in amino acids and small peptides. (Adapted from G. N. Ramachandran, et al., Biochim. Biophys. Acta 359:298–302, 1974.)

The presence of an asymmetric center at the C^{α} carbon atom, with only L-amino acid residues, results in an inherent asymmetry of the polypeptide chain. This will be shown to be important for its spectral properties and conformation. In principle, rotation may occur about any of the three bonds of the polypeptide backbone. However, the peptide bond appears to have partial double-bonded character, presumably owing to resonance:



(1–7)

so that the six atoms depicted above have a strong tendency to be coplanar. Consequently, the C′—N bond length is observed crystallographically to be 1.33 Å, shorter than a normal bond length of 1.45 Å, as in the C^{α}—N bond, but also longer than the value of 1.25 A for the average C═N bond length in model compounds. The peptide bond appears to have approximately 40 percent double-bonded character. Rotation of this bond is then markedly restricted, but two configurations of the planar peptide bond are possible, one in which the C^{α} atoms are *trans*, the other with them *cis:*



(1–8)

The *trans* form is favored energetically, probably owing to fewer repulsions between nonbonded atoms.

As a result of this resonance, the peptide backbone is relatively polar, with the electrons tending to be redistributed to give the permanent dipole indicated in the resonance state depicted in Equation 1–7. The H and N atoms appear to have positive and negative, respectively, equivalent charges of 0.28 electron, while C and O have respective positive and negative equivalent charges of 0.39 electron.

The other two types of bonds of the peptide backbone appear to behave as normal C—C and C—N bonds.

The peptide backbone is chemically rather unreactive. A proton is added or lost only at extremes of pH. The apparent pK_a value for deprotonation of the amide NH of the polypeptide backbone is between 15 and 18, whereas it is about -8 to -12 for protonation. The oxygen atom is protonated more readily, with an apparent pK_a value of about -1. These properties are extremely useful for making possible measurement of hydrogen exchange from the backbone, which is important for studying the conformation in solution (see Chapter 7). At pH values where protonation or deprotonation becomes significant, the polypeptide chain is generally hydrolyzed to yield the substituent amino acids. For example, the standard method for determining the amino acid composition of proteins or peptides is to heat them at 105° C for about 24 hours in 6 M HCl. Other chemical alterations of the peptide chain require either drastic conditions or the relatively close proximity of certain reactive groups attached to the amino acid side chains, which may be used profitably for selective cleavage of the polypeptide chain at certain amino acid residues (see later section, Fragmentation of a Protein into Peptides).

The Nature of the Chemical Bond, 3rd ed. L. Pauling. Ithaca, N.Y., Cornell University Press, 1960.

Conformational energy estimates for statistically coiling polypeptide chains. D. A. Brant, et al. J. Mol. Biol. 23:47–65, 1967.

Abbreviations and symbols for the description of the conformation of polypeptide chains. IUPAC–IUB Commission on Biochemical Nomenclature. J. Mol. Biol. 52:1–17, 1970.

Hydrogen exchange. S. W. Englander, et al. Ann. Rev. Biochem. 41:903–924, 1972.

Coordinating properties of the amide bond. Stability and structure of metal ion complexes of peptides and related ligands. H. Sigel and R. B. Martin. Chem. Rev. 82:385–426, 1982.

AMINO ACID SIDE CHAINS

The 20 different amino acid side chains possess a variety of chemical properties which, when combined on a single molecule, give a protein properties far beyond those possible with organic molecules. This chemical diversity is vital for the unique functions of biological proteins. As will become apparent, the chemical properties of the whole protein are far greater than the sum of its constituent amino acids, but the intrinsic chemical properties of these 20 different side chains are important as a starting point for considering their roles in proteins.

Some pertinent chemical and physical properties of the 20 amino acid residues are summarized in Table 1–1. References to their detailed geometries determined crystallographically are given in Table 1–2. Each of the side chains will be discussed briefly, describing their normal chemical properties. However, specific residues in folded proteins may have very different properties, as will be discussed in Chapter 7.

The reactions by which the various side chains can be modified will be those that are moderately selective and specific, so that they may be used on proteins and peptides.

Chemistry of the Amino Acids. J. P. Greenstein and M. Winitz. New York, John Wiley & Sons, 1961.

Chemical Modification of Proteins. G. E. Means and R. E. Feeney. San Francisco, Holden-Day, 1971.

Table 1–1 Properties of Amino Acid Residues

	ONE-LETTER ABBREVIATION	MASS[a] (daltons)	VOLUME[b] (Å^3)	ACCESSIBLE SURFACE AREA[c] (Å^2)	PARTIAL SPECIFIC VOLUME[b] (ml/gm)	pK$_a$ OF IONIZING SIDE CHAIN[d]	OCCURRENCE IN PROTEINS[e] (%)
Ala	A	71.08	88.6	115	0.748		9.0
Arg	R	156.20	173.4	225	0.666	~12	4.7
Asn	N	114.11	117.7	160	0.619		4.4
Asp	D	115.09	111.1	150	0.579	4.5	5.5
Cys	C	103.14	108.5	135	0.631	9.1 to 9.5	2.8
Gln	Q	128.14	143.9	180	0.674		3.9
Glu	E	129.12	138.4	190	0.643	4.6	6.2
Gly	G	57.06	60.1	75	0.632		7.5
His	H	137.15	153.2	195	0.670	6.2	2.1
Ile	I	113.17	166.7	175	0.884		4.6
Leu	L	113.17	166.7	170	0.884		7.5
Lys	K	128.18	168.6	200	0.789	10.4	7.0
Met	M	131.21	162.9	185	0.745		1.7
Phe	F	147.18	189.9	210	0.774		3.5
Pro	P	97.12	122.7	145	0.758		4.6
Ser	S	87.08	89.0	115	0.613		7.1
Thr	T	101.11	116.1	140	0.689		6.0
Trp	W	186.21	227.8	255	0.734		1.1
Tyr	Y	163.18	193.6	230	0.712	9.7	3.5
Val	V	99.14	140.0	155	0.847		6.9
						α-Amino 6.8 to 7.9	
						α-Carboxyl 3.5 to 4.3	

[a]Molecular weight of amino acid minus that of water. Values from Handbook of Chemistry and Physics, 43rd ed. Cleveland, Chemical Rubber Publishing Co., 1961.

[b]Values from A. A. Zamyatnin, Prog. Biophys. Mol. Biol. 24:107–123, 1972.

[c]Values from C. Chothia, J. Mol. Biol. 105:1–14, 1975. The accessible surface area is defined in Figure 6–20.

[d]Estimated from the pK$_a$ values of small model compounds, from C. Tanford, Adv. Protein Chem. 17:69–165, 1962.

[e]Frequency of occurrence of each amino acid residue in the primary structures of 207 unrelated proteins of known sequence. Values from M. H. Klapper, Biochem. Biophys. Res. Commun. 78:1018–1024, 1977.

Table 1–2 Reference Sources for Detailed Geometries of the Amino Acids

Alanine	M. S. Lehmann, et al., J. Am. Chem. Soc. 94:2657–2663, 1972
Arginine	M. S. Lehmann, et al., J. Chem. Soc. [Perkin II] 2:133–135, 1973
Asparagine	J. J. Verbist, et al., Acta Cryst. B28:3006–3013, 1972
Aspartic acid	D. S. Eggleston, et al., Acta Cryst. B37:1428–1430, 1981
Cysteine	K. A. Kerr and J. P. Ashmore, Acta Cryst. B29:2124–2127, 1973
Cystine	D. D. Jones, et al., Acta Cryst. B30:1220–1228, 1974
Glutamic acid	M. S. Lehmann, et al., J. Cryst. Mol. Struct. 2:225–233, 1972
Glutamine	T. F. Koetzle, et al., Acta Cryst. B29:2571–2575, 1973
Glycine	D. S. Eggleston, et al., Acta Cryst. B37:1427–1430, 1981
Histidine	T. J. Kristenmacher, et al., Acta Cryst. B28:3352–3361, 1972
Isoleucine	K. Torii and Y. Iitaka, Acta Cryst. B27:2237–2246, 1971
Leucine	L. Golic and W. C. Hamilton, Acta Cryst. B28:1265–1271, 1972
Lysine	T. F. Koetzle, et al., Acta Cryst. B28:3207–3214, 1972
Methionine	C. Chen and R. Parthasarathy, Acta Cryst. B33:3332–3336, 1977
Phenylalanine	A. R. Al-KaraGhouli and T. F. Koetzle, Acta Cryst. B31:2461–2465, 1975
Proline	I. Tanaka, et al., Acta Cryst. B33:116–119, 1977
Serine	M. N. Frey, et al., Acta Cryst. B29:876–884, 1973
Threonine	V. S. Yadava and V. M. Padmanabhan, Acta Cryst. B29:854–858, 1973
Tryptophan	L. C. Andrews, et al., Amer. Cryst. Assoc. Abstracts G6, 1974
Tyrosine	M. N. Frey, et al., J. Chem. Phys. 56:2547–2556, 1973
Valine	T. F. Koetzle, et al., J. Chem. Phys. 60:4690–4696, 1974

The chemical modification of proteins by group-specific and site-specific reagents. A. N. Glazer. In The Proteins, 3rd ed. H. Neurath and R. L. Hill (eds.). Vol. 2, pp. 1–103. New York, Academic Press, 1976.

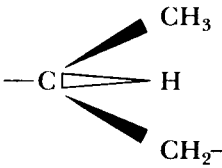
Electronic structure and bonding of the amino acids containing first row atoms. D. A. Dixon and W. N. Lipscomb. J. Biol. Chem. 251:5992–6000, 1976.

Gly

Gly is the simplest amino acid, with no side chain. Because there are two H atoms on C^{α}, it is not asymmetric and is the only amino acid that cannot exist as D or L isomers. The absence of a side chain gives the polypeptide backbone at Gly residues much greater conformational flexibility than that possible otherwise, as will be discussed further in Chapter 5.

The Aliphatic Residues: Ala, Val, Leu, Ile

With solely aliphatic side chains, containing no polar or functional chemical groups, Ala, Val, Leu, and Ile comprise a rather homogeneous class. Their inert side chains have the chemical property of being hydrophobic; thus they have the potential for being structural units, since they provide a variety of molecular surfaces and shapes that seem to be well-suited for such a purpose. It might be noted that there occurs among the normal 20 amino acids used in proteins none with a single ethyl side chain, corresponding to that of the amino acid α-amino butyric acid. The side chain of isoleucine has an extra center of asymmetry; only the one isomer



(1–9)

occurs naturally and is incorporated into proteins.