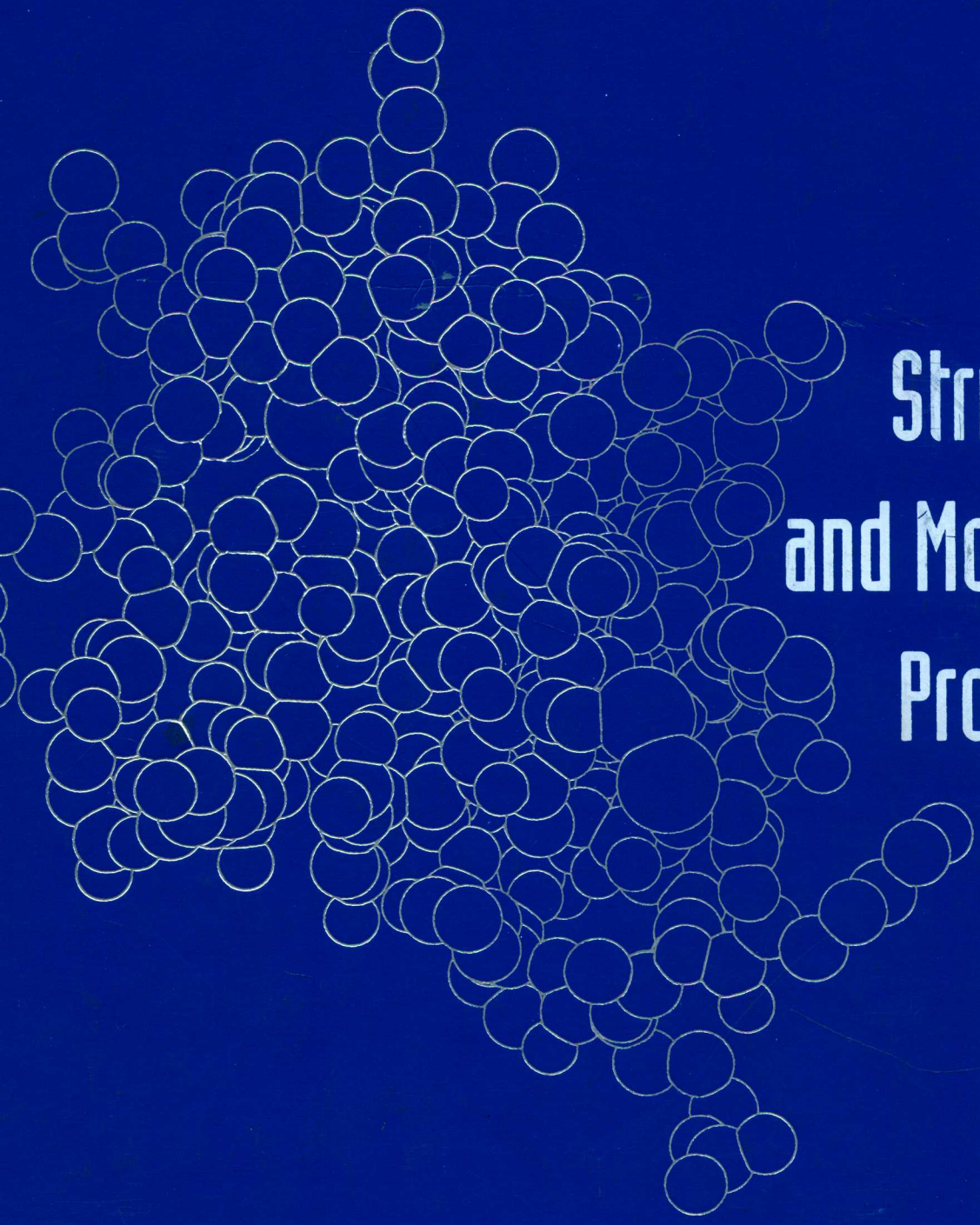
SECOND EDITION

PROTEINS



Structures and Molecular Properties

THOMAS E. CREIGHTON

PROTEINS

Structures and Molecular Properties

Second Edition

Thomas E. Creighton
European Molecular Biology Laboratory
Heidelberg, Germany

W. H. Freeman and Company • New York

Cover image provided by Martin Noble.

Library of Congress Cataloging-in-Publication Data
Creighton, Thomas E., 1940–
Proteins : structures and molecular properties / Thomas E. Creighton. — 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-7167-7030-X
1. Proteins — Structure. 2. Proteins — Chemistry. I. Title.
QP551.C737 1993
574.19′245 — dc20 92-6664
CIP

Printed in the United States of America

Sixth printing, 2002

PROTEINS

Second Edition

Preface

Because our understanding of protein structure and function has increased remarkably in the nine years since the first edition of this volume, most of this edition needed to be entirely rewritten. The structures of the twenty amino acid residues used in proteins, at least, have not changed. Also, there has been no marked change in our understanding of their intrinsic chemical properties (Chapter 1), although this probably reflects the decreased importance of chemical modification for studying protein function (Sec. 1.3) due to the advent of site-directed mutagenesis and protein engineering (Sec. 2.2). Procedures for the analysis, sequencing, and synthesis of peptides and proteins (Secs. 1.6 and 1.8) have been improved and have become routine and automated. A thorough knowledge of the chemical basis of such techniques is no longer absolutely necessary, but any serious scientist needs to understand them so as not to be misled by results obtained with automated procedures. The classical technique of mass spectrometry previously seemed of only limited use with nonvolatile samples like proteins, but new developments have made it of major importance in studying protein structure (Sec. 1.6.2.d).

Many new posttranslational modifications of proteins have been found, a number of physiological importance, and those best characterized are described in Section 2.4. The twenty-first amino acid to be incorporated directly into proteins during biosynthesis, selenocysteine, has been discovered (Sec. 2.1.3.d). The overall process of protein biosynthesis is described in Section 2.1.

One of the greatest technological advances has been the introduction of protein engineering (Sec. 2.2), whereby a protein of any desired amino acid sequence can be produced in substantial quantities by expressing a cloned or synthetic gene. The primary structure can be altered with remarkable specificity

by site-directed mutagenesis of the gene. This has had a major impact on the study of protein structure, stability, and function, as described in Chapters 6 through 9. The relative simplicity of gene cloning and sequencing has resulted in an explosion of primary structures of proteins determined in this way, although the variety of posttranslational modifications possible requires the covalent structure of the final active protein to be determined. Even proteins that occur naturally in only minute quantities, and therefore are often those of greatest biological potency, can be identified and sequenced from their genes. Such proteins can then be produced in quantities sufficient for industrial or pharmaceutical use by expressing these genes.

The ease of sequencing proteins via their genes has resulted in great pressure to obtain further insight into the function of a protein from just its primary structure. This is best accomplished by scanning the sequence data banks (Appendix 1) for homologous proteins of known function or structure. It is too frequently assumed, however, that homologous proteins share the same functions, and an understanding of the relationship between protein structure and function, as explained in Chapters 6 through 9, is necessary to minimize the many opportunities to be misled.

After a protein polypeptide chain is synthesized, it must also be directed to its appropriate place in the cell, and great advances have been made in this area (Sec. 2.3). The roles of the very many proteins that are present in a cell continue to be studied in increasing numbers, and the cell biology of proteins will become increasingly important. As larger and more complex proteins are studied, the challenge of extrapolating our understanding of the fundamental properties of simpler proteins to the more complex cases will become greater. In many cases, the fundamental knowledge is available but overlooked, as in the rediscovery of coiled coils (Sec. 5.5.2) as so-called leucine zippers.

One striking aspect of the study of protein structure is how limited our understanding is of the fundamental physical chemistry that is the basis for protein structure and function (Chapter 4). For example, our understanding of the hydrophobic interaction has changed substantially in the past few years (Sec. 4.3). How to analyze the simplest physical interaction, that between charged groups, in complex systems like proteins, is only now becoming clear. Entropic considerations for protein structure and function (Sec. 4.4) are still largely ignored, and it is possible to calculate free-energy changes in proteins only when very small changes take place (Sec. 7.4). Although much has been accomplished, much remains to be learned or applied.

The technique of X-ray crystallography has been developed and refined further (Sec. 6.1), and crystallographers determine increasing numbers of protein structures, each more beautiful, awesome, and accurate than the last. There are now approximately 500 entries in the list of known protein structures (Appendix 2), whereas the first edition listed only 148. A major advance in determining protein structure has been the development of nuclear magnetic resonance spectroscopy (NMR) as a tool for determining the polypeptide chain fold in small proteins (Sec. 6.3). With technical advances and isotopic labeling, larger and larger proteins are being studied in this way. NMR has the advantages of allowing proteins to be studied in solution, thus eliminating the need for crystalline samples, and of giving dynamic information, but X-ray crystallography is also becoming a much more rapid and dynamic technique (Sec. 6.1.7).

With the availability of many more protein structures, the rules of protein architecture are slowly becoming understood. The power of protein engineering techniques for studying protein structure (Sec. 6.5), stability (Sec. 7.4), folding (Sec. 7.5), and function (Chapters 8 and 9) has resulted in a great deal of experimental data, but relatively few simple conclusions. Progress in being able to predict protein structure and function from just the amino acid sequence is painfully slow and is predominantly an empirical process, due largely to the complexity of the problem. The protein folding problem may be solved when we know most of the possible protein tertiary structures, which may number no more than 500 to 1000, so that every protein is found from its amino acid sequence to be homologous to a protein of known structure and function.

Membrane proteins were largely ignored in the first edition, but the high-resolution structures of the bacterial photosynthetic reaction center and porin have made it possible to describe this important class of proteins in molecular terms (Sec. 7.2). These two membrane proteins have turned out to have structures remarkably like water-soluble proteins but with nonpolar surfaces where they are interacting with the nonpolar parts of the lipid bilayer. The interactions of globular proteins with the aqueous solvent are now understood to a great extent, and it is now possible to discuss the aqueous solubilities of proteins in considerable detail (Sec. 7.1).

Virtually all proteins function by interacting specifically with other molecules (Chapter 8), and the greatest advance in this area has been with DNA-binding proteins (Sec. 8.3.2). These proteins play crucial roles in the replication and expression of genetic information and are central to understanding cell function. It is now clear that a number of protein structural motifs and types of interaction with nucleic acids are involved, but there is no simple code relating the sequence of the

protein to the sequence of the DNA that it recognizes. Consequently, a comprehensive understanding is required of all the various types of DNA-binding proteins, so that all the many proteins involved in gene regulation and expression can be identified and characterized from just their amino acid sequences.

The interactions of antibodies with antigens have been characterized in detail in a number of instances, including cases where the antigen is another protein molecule (Sec. 8.3.1). This interaction was predicted to have some special properties, but all the observations so far indicate that it is typical of other protein–ligand interactions, described in Chapter 8. The phenomenon of allostery involving positive cooperativity of binding is now fairly clear (Secs. 8.4 and 9.4.2), but that concerning negative cooperativity largely remains a mystery (Sec. 8.4.5).

The study of enzyme catalysis has been revolutionized by the technology of protein engineering (Chapter 9), and it is now possible to measure quantitatively the contribution to binding and catalysis of each group of the side chains of the enzyme. Consequently, very detailed enzyme mechanisms are now known in a few instances (Sec. 9.3). It is clear that the most important property of enzymes is their complementarity to the transition state of the reaction (Sec. 9.2). This probably explains why increased binding interactions between a substrate and enzyme are so often manifested as increases in catalytic rate rather than just binding affinity, although a detailed explanation is not yet possible. The importance of complementarity to the transition state is also demonstrated by the generation of catalytic antibodies using transition-state analogues as immunogens (Sec. 9.2.4).

The importance of enzyme regulation by reversible covalent modification, especially phosphorylation, is increasingly evident, and the molecular basis of this phenomenon is now known in detail in two instances, glycogen phosphorylase and isocitrate dehydrogenase (Sec. 9.4.3).

Finally, the complex phenomenon of protein degradation is now known to be of importance physiologically, and the mechanism of the process is understood to a remarkable extent. Little was known nine years ago.

With all this new knowledge, the description of proteins presented here is much more complete than was possible with the first edition. As a consequence, the volume has grown somewhat larger, although this was kept to a minimum. Including the necessary recent references while minimizing the book's length required that many of the earlier references be omitted from this edition. It is recommended, therefore, that the first edition also be consulted for a more complete listing of references. I hope that the second edition has fewer errors than the first, and I wish to thank those who pointed out deficiencies and errors in the first edition. Various sections of this edition have been read by some of those most active in the field, and I would like to thank especially R. L. Baldwin, P. R. Evans, J. Ewbank, A. R. Fersht, L. N. Johnson, J. R. Knowles, W. Lipscomb, B. W. Matthews, E. Meyer, and H. K. Schachman for their valuable assistance. Of course, perfection would have required an undue delay in publication, and I remain responsible for any errors that might remain.

An innovation in this edition is the exercises at the end of each chapter. They are intended as informative and instructive exercises by which students can expand their understanding of proteins and of how scientific research progresses, primarily by referring to the literature. Most of the examples are of instances where errors or alternative interpretations were reported in the literature, because it is felt that such instances illustrate most clearly the adventure of scientific research. Most papers in the literature are highly polished, condensed, and edited reports of a selected aspect of a scientific investigation and do not report the often convoluted and erratic process by which the end product was attained. One can learn from the literature how science progresses in this way only when such instances are brought out into the open. It is in this spirit that the exercises are offered as interesting, informative, and sometimes amusing glimpses into how proteins are studied.

Of course, there is no substitute for participating in the adventure, and it is hoped that this volume may inspire young people with the appropriate interest and motivation to contribute to future research on the fascinating subject of proteins.

Thomas E. Creighton
May 1992

PROTEINS

Second Edition

Contents

1

Chemical Properties of Polypeptides

Virtually every property that characterizes a living organism is affected by proteins. Nucleic acids, also essential for life, encode genetic information —mostly specifications for the structures of proteins—and the expression of that information depends almost entirely on proteins (though some RNA molecules with catalytic activity have been discovered recently).

Life forms make use of many chemical reactions to supply themselves continually with chemical energy and to use it efficiently, but by themselves these reactions could not occur fast enough under physiological conditions (aqueous solution, 37 °C, pH 7, atmospheric pressure) to sustain life. The rates of these reactions are increased, by many orders of magnitude, in organisms by the presence of enzymes, which also are proteins. The subject of biochemistry is primarily a study of the roles of enzymes in living systems.

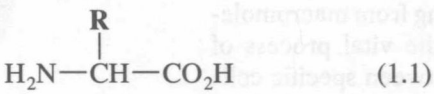
Proteins store and transport a variety of particles ranging from macromolecules to electrons. They guide the flow of electrons in the vital process of photosynthesis; as hormones, they transmit information between specific cells and organs in complex organisms; some proteins control the passage of molecules across the membranes that compartmentalize cells and organelles; proteins function in the immune systems of complex organisms to defend against intruders (the best known are the antibodies); and proteins control gene expression by binding to specific sequences of nucleic acids, thereby turning genes on and off. Proteins are the crucial components of muscles and other systems for converting chemical energy into mechanical energy. They also are necessary for sight, hearing, and the other senses. And many proteins are simply structural, providing the filamentous architecture within cells and the materials that are used in hair, nails, tendons, and bones of animals.

In spite of these diverse biological functions, proteins are a relatively homogeneous class of molecules. All are the same type of linear polymer, built of various combinations of the same 20 amino acids. They differ only in the sequence in which the amino acids are assembled into polymeric chains. The secret to their functional diversity lies partly in the chemical diversity of the amino acids but primarily in the diversity of the three-dimensional structures that these building blocks can form, simply by being linked in different sequences. The awesome functional properties of proteins can be understood only in terms of their relationship to the three-dimensional structures of proteins. That relationship is the topic of this volume.

1.1 The Polymeric Nature of Proteins

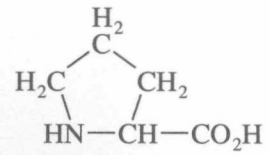
Proteins are more complex than most linear polymers in that they can incorporate 20 different monomers in their construction instead of only one or two. In another sense, proteins are structurally less complex. Most chemical polymers are synthesized by chemically polymerizing a mixture of monomers, thereby producing a distribution of chain lengths and, if more than one type of monomer is present, an approximately random sequence of monomers. Natural proteins, however, are linear and unbranched and have precise lengths and exact sequences of amino acids. Indeed, it is only the differences in length and sequence that distinguish one protein from any other and that make possible a diversity of structures and functions. Most importantly, the linear polymeric chain of almost every natural protein has the crucial property of being able to assume a specific three-dimensional folded conformation, as will be described in later chapters.

Of the 20 amino acids usually found in proteins, 19 have the general structure



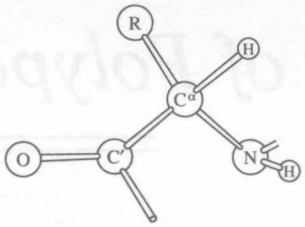
(1.1)

and differ only in the chemical structures of the side chain **R**. The 20th natural amino acid, proline, is similar, but its side chain is bonded to the nitrogen atom to give the imino acid:



(1.2)

Except in glycine, where the side chain is only a hydrogen atom, the central carbon atom is asymmetric and is always the L isomer:

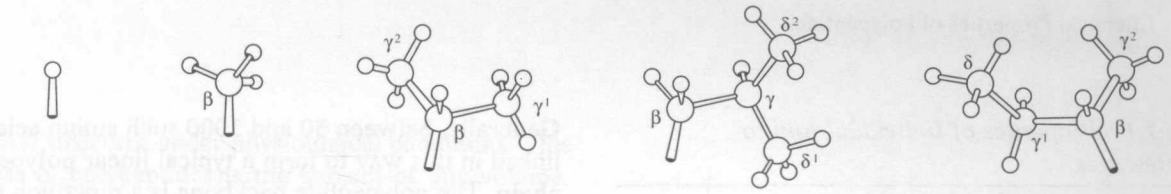


(1.3)

Unless indicated otherwise, all amino acids in this volume are L isomers.

The structures of the side chains of the 20 amino acids are illustrated in Figure 1.1. The central carbon atom depicted in Equation (1.1) is designated as α, and the atoms of the side chains are commonly designated β, γ, δ, ϵ, and ζ, in order away from the α carbon atom. Chemical groups, however, are usually designated by the carbon atom to which they are bonded; hence, the ζN atom of a Lys residue is part of the ϵ-amino group. Table 1.1 and Figure 1.1 also give the three- and one-letter abbreviations commonly used for designating the amino acids when they are incorporated into proteins. Three-letter abbreviations are used throughout this volume because their designations are obvious. The relationships of the one-letter codes to the amino acids are somewhat less obvious, but they are commonly used in compilations of long sequences because they save space and are less likely to be confused (e.g., Gln, Glu, and Gly can easily be confused but not Q, E, and G).

FIGURE 1.1
Side chains of the 20 amino acids that occur naturally in proteins. Small unlabeled spheres are hydrogen atoms, and large unlabeled spheres are carbon atoms; other atoms are labeled. Double bonds are black, and partial double bonds are shaded. In the case of Pro, the bonds of the polypeptide backbone are included and are black. Below the name of the amino acid are the three-letter and the one-letter abbreviations commonly used. Note that isoleucine and threonine have asymmetric centers in their side chains, and only the isomer illustrated is used biologically.



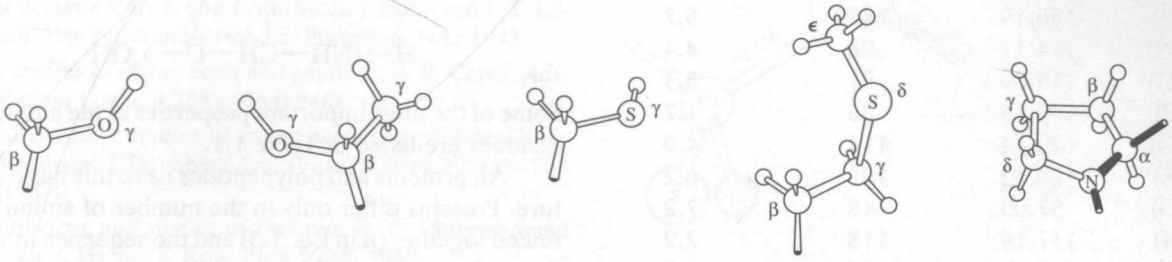
Glycine
Gly
G

Alanine
Ala
A

Valine
Val
V

Leucine
Leu
L

Isoleucine
Ile
I



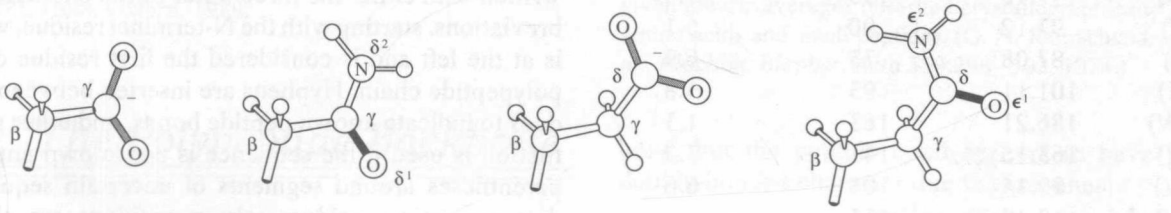
Serine
Ser
S

Threonine
Thr
T

Cysteine
Cys
C

Methionine
Met
M

Proline
Pro
P

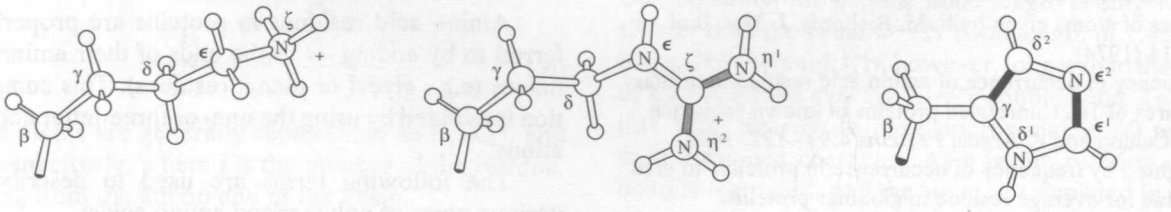


Aspartic acid
Asp
D

Asparagine
Asn
N

Glutamic acid
Glu
E

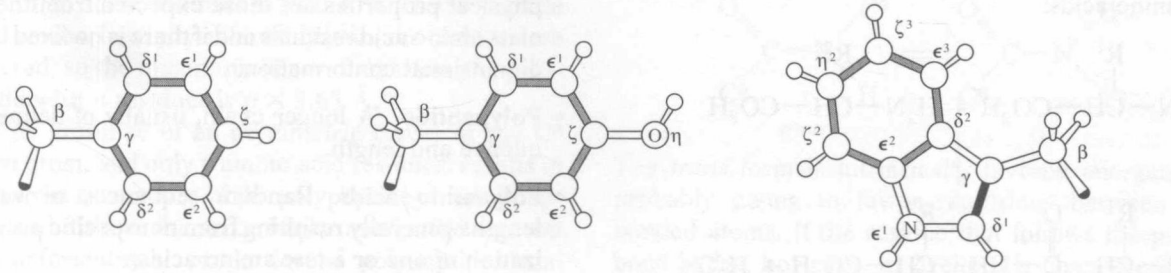
Glutamine
Gln
Q



Lysine
Lys
K

Arginine
Arg
R

Histidine
His
H



Phenylalanine
Phe
F

Tyrosine
Tyr
Y

Tryptophan
Trp
W