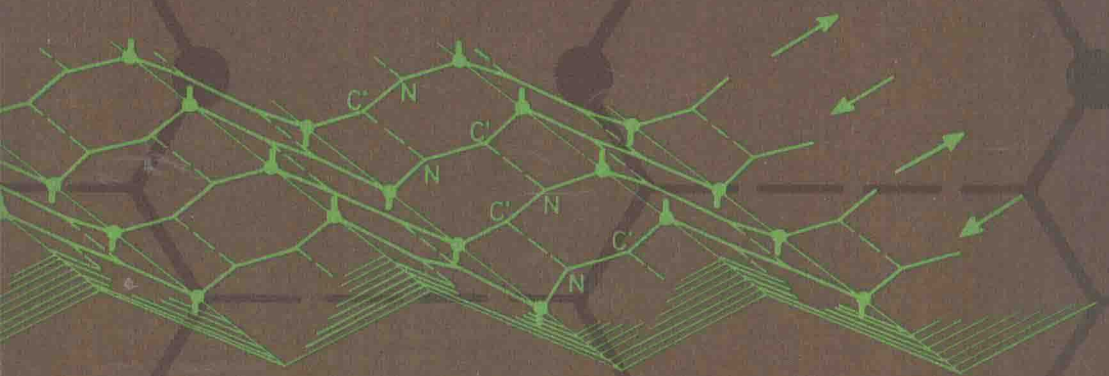


G. E. Schulz
R. H. Schirmer

Principles of Protein Structure



Springer-Verlag
New York Heidelberg Berlin

G. E. Schulz
R. H. Schirmer

Principles of Protein Structure



Springer-Verlag
New York Heidelberg Berlin

Dr. Georg E. Schulz
Max-Planck-Institut für Medizinische Forschung
Jahnstr. 29
6900 Heidelberg
Federal Republic of Germany

Dr. R. Heiner Schirmer
Max-Planck-Institut für Medizinische Forschung
Jahnstr. 29
6900 Heidelberg
Federal Republic of Germany

Series Editor:

Prof. Charles R. Cantor
Columbia University
Box 608 Havemeyer Hall
New York, New York 10027 USA

Library of Congress Cataloging in Publication Data

Schulz, Georg E., 1939-
Principles of protein structure.
(Springer advanced texts in chemistry)
Bibliography: p.
Includes index.
I. Proteins. I. Schirmer, R. Heiner,
1942- joint author. II. Title. III. Series.
QP551.S4255 547'.75 78-11500

All rights reserved.

No part of this book may be translated or reproduced in any form without written permission
from Springer-Verlag.

© 1979 by Springer-Verlag New York Inc.

Printed in the United States of America

9 8 7 6 5 4 3 2 1

Paper Edition:

ISBN 0-387-90334-8 Springer-Verlag New York
ISBN 3-540-90334-8 Springer-Verlag Berlin Heidelberg

Cloth Edition:

ISBN 0-387-90386-0 Springer-Verlag New York
ISBN 3-540-90386-0 Springer-Verlag Berlin Heidelberg

Springer Advanced Texts in Chemistry

Charles R. Cantor, Editor

Series Preface

New textbooks at all levels of chemistry appear with great regularity. Some fields like basic biochemistry, organic reaction mechanisms, and chemical thermodynamics are well represented by many excellent texts, and new or revised editions are published sufficiently often to keep up with progress in research. However, some areas of chemistry, especially many of those taught at the graduate level, suffer from a real lack of up-to-date textbooks. The most serious needs occur in fields that are rapidly changing. Textbooks in these subjects usually have to be written by scientists actually involved in the research which is advancing the field. It is not often easy to persuade such individuals to set time aside to help spread the knowledge they have accumulated. Our goal, in this series, is to pinpoint areas of chemistry where recent progress has outpaced what is covered in any available textbooks, and then seek out and persuade experts in these fields to produce relatively concise but instructive introductions to their fields. These should serve the needs of one semester or one quarter graduate courses in chemistry and biochemistry. In some cases the availability of texts in active research areas should help stimulate the creation of new courses.

New York, New York
December, 1978

CHARLES R. CANTOR

Preface

Among the four major types of biopolymers—nucleic acids, polysaccharides, lipid assemblies, and proteins—the latter are best understood because they are structurally well defined and therefore suitable for isolation and analysis. Proteins were placed at the forefront of biopolymer research more than a century ago when Kühne (1) purified and characterized the enzyme trypsin and Hoppe-Seyler (2) obtained crystals of haemoglobin. These crystals indicated that proteins were uniform down to atomic detail; but this fact was not widely recognized until atomic detail was actually made visible by X-ray diffraction techniques.

Today more than 70 protein structures are known. From a practical viewpoint these structures provide biochemists and pharmacologists with models for the design and interpretation of experiments. In general, they put biochemical sciences on a sounder basis. The aim of this book is to consider principles which have emerged from comparisons of known protein structures. These principles are the first step for understanding the relationship between a protein's blueprint as stored in DNA and the action of this protein in an organism. In addition, establishing principles facilitates and extends the usage of the complex protein architectures for scientific and practical purposes. The very existence of structural principles indicates that the diversity of protein functions is not paralleled by structural diversity. For the biomedical sciences this could mean that a comprehensive understanding of physiological and pathological processes at the molecular level is not a utopian dream, and that it may become possible to influence such processes in pin-pointed fashion.

Each chapter of this book is self-contained. The chapters can be read in any order, but those unfamiliar with protein structures are recommended to read chapter seven first. We would like to thank all our colleagues who allowed us to use their graphic and digital material. Furthermore, we thank all those who helped us in preparing the manuscript.

Principles of Protein Structure

Contents

Chapter 1

Amino Acids	1
1.1 The 20 Standard Amino Acids	1
1.2 Why Were Just These Amino Acids Selected?	1
1.3 Colinear Relation Between Nucleic Acids and Polypeptides	9
1.4 Side Chain Properties	10
1.5 Empirical Similarities Between Amino Acid Residues	14
Summary	16

Chapter 2

Structural Implications of the Peptide Bond	17
2.1 Synthesis on Ribosomes	17
2.2 Peptide Bond Dimensions	18
2.3 Steric Hindrance	20
2.4 Conformational Energy	21
2.5 <i>cis</i> versus <i>trans</i> Configuration	25
Summary	26

Chapter 3

Noncovalent Forces Determining Protein Structure	27
3.1 Dispersion Forces and Electron Shell Repulsion	27
3.2 Electrostatic Interactions	30

3.3	Van der Waals Potentials	32
3.4	Hydrogen Bonds	33
3.5	Entropic Forces	36
3.6	Molecular Packing	42
	Summary	44
Chapter 4		
	The Covalent Structure of Proteins	46
4.1	Chain Assemblies	46
4.2	Disulfide Bonds	53
4.3	Enzyme-Controlled Modifications of the Main Chain	56
4.4	Enzyme-Controlled Side Chain Modifications	62
	Summary	65
Chapter 5		
	Patterns of Folding and Association of Polypeptide Chains	66
5.1	Secondary Structure	66
5.2	Supersecondary Structures	79
5.3	Structural Domains	84
5.4	Globular Proteins	95
5.5	Aggregates of Globular Proteins	98
5.6	Hierarchy of Levels	105
	Summary	106
Chapter 6		
	Prediction of Secondary Structure from the Amino Acid Sequence	108
6.1	Probabilistic Methods	110
6.2	Physico-Chemical Methods	114
6.3	Application of Prediction Methods	120
6.4	Evaluation of Prediction Methods	122
6.5	Emerging General Results	128
	Summary	129
Chapter 7		
	Models, Display, and Documentation of Protein Structures	131
7.1	Structural Ingredients	131
7.2	Representations of Complete Structures	134
7.3	Abstract Chain Fold Representations	144
	Summary	148

Chapter 8

Thermodynamics and Kinetics of Polypeptide Chain Folding 149

8.1 Thermodynamic Aspects	149
8.2 Speed, Precision, and Limitations of Folding <i>in vitro</i>	153
8.3 Structural Elements in Unfolded Chains	155
8.4 Folding Pathway	157
8.5 The Influence of Ligands	160
8.6 Simulations of the Folding Process	162
Summary	164

Chapter 9

Protein Evolution 166

9.1 Protein Speciation	168
9.2 Phylogeny on the Basis of Protein Structures	175
9.3 Protein Differentiation	179
9.4 Gene Fusion	193
9.5 Convergent Evolution in Proteins	197
9.6 Recognition of Distant Relationships	198
Summary	205

Chapter 10

Protein–Ligand Interactions 206

10.1 Ligand-Binding Sites of Immunoglobulins	206
10.2 Substrate-Binding Sites of Serine Proteases	209
10.3 Haem-Binding Sites	211
10.4 Nucleotide-Binding Sites	220
10.5 Binding Sites for Phosphoryl Groups	224
10.6 Interactions of Proteins With Other Macromolecules	227
Summary	231

Chapter 11

The Structural Basis of Protein Mechanism, Action, and Function 233

11.1 Definitions	233
11.2 Enzyme Catalysis	234
11.3 Biological and Medical Aspects of Protein Action and Function	241
11.4 Skeletal Muscle, a System in Which Protein Action Can be Related to the Overall Performance of an Organ	243
Summary	250

Appendix	
Statistical Mechanics of the Helix-Coil Transition	252
A.1 Partition Function	252
A.2 Probability of a Residue Being in a Certain Conformation	252
A.3 Ising Model	254
A.4 Zimm-Bragg Model for Helix-Coil Transition	254
A.5 Comparison With Experimental Data	260
References	262
Index	303

Chapter 1

Amino Acids

1.1 The 20 Standard Amino Acids

Today, the basic organizational scheme of living material is known. Nucleic acids contain all genetic information, which is laid down in a sequence of four different nucleotide bases along a chain. There are two levels of nucleic acids. The more stable deoxyribonucleic acid (DNA) is the permanent information store. The less stable ribonucleic acid (RNA), which is “transcribed” from DNA, comes as a handy blueprint that is “translated” into the amino acid sequence of proteins using the ribosomal machinery. Proteins perform virtually all types of work in an organism.

In this chapter we are concerned with the 20 standard amino acid residues occurring in proteins. These are listed in Table 1-1 according to their relative abundance. The commonly used three-letter and one-letter abbreviations together with mnemonic aids are given. The molecular weights of the residues range from 57 to 186 daltons, the weighted average being about 110. Thus, a protein with a molecular weight of 33,000 contains approximately 300 residues. The amino acids can be classified with respect to their side chains as either polar or nonpolar. The polar side chains can be further subdivided into neutral, basic, or acidic. In Figure 1-1, which illustrates the 20 common amino acid side chains, residues with similar properties are grouped near one another.

1.2 Why Were Just These Amino Acids Selected?

Presumably, life started with autocatalytic nucleic acids and not with proteins. The question of why just these amino acids were selected cannot yet be answered. Nevertheless speculating about it can provide us with some insights. First, let us discuss the origin of the basic organizational scheme introduced above, the origin of life. It is highly probable that life

Table 1-1. Notations and Properties of the 20 Standard Amino Acid Residues of Proteins^a

Amino acid or residue thereof	Three-letter symbol	One letter symbol	Mnemonic help for one- letter symbol	Relative abundance in <i>E. coli</i> proteins (19) (%)	M.W. of residue at pH7.0 (daltons)	pK value of side chain (19)	ΔG values for transfer of side chain from water to ethanol at 25°C (16) (kcal/mol)
Alanine	Ala	A	<u>A</u> lanine	13.0	71		-0.5
Glutamate	Glu	E	glu <u>E</u> ttamic acid		128	4.3	
Glutamine	Gln	Q	<u>Q</u> -tamine	10.8	128		
Aspartate	Asp	D	aspar <u>D</u> ic acid	9.9	114	3.9	
Asparagine	Asn	N	aspar <u>D</u> ic <u>N</u> e		114		
Leucine	Leu	L	<u>L</u> eucine	7.8	113		-1.8
Glycine	Gly	G	<u>G</u> lycine	7.8	57		
Lysine	Lys	K	before <u>L</u> e	7.0	129	10.5	
Serine	Ser	S	<u>S</u> erine	6.0	87		+0.3
Valine	Val	V	<u>V</u> aline	6.0	99		-1.5
Arginine	Arg	R	a <u>R</u> gine	5.3	157	12.5	
Threonine	Thr	T	<u>T</u> hreonine	4.6	101		-0.4
Proline	Pro	P	Proline	4.6	97		
Isoleucine	Ile	I	<u>I</u> soleucine	4.4	113		
Methionine	Met	M	<u>M</u> ethionine	3.8	131		-1.3
Phenylalanine	Phe	F	Phenylalanine	3.3	147		-2.5
Tyrosine	Tyr	Y	t <u>Y</u> rosine	2.2	163	10.1	-2.3
Cysteine	Cys	C	<u>C</u> ysteine	1.8	103		
Tryptophan	Trp	W	t <u>W</u> o rings	1.0	186		-3.4
Histidine	His	H	<u>H</u> istidine	0.7	137	6.0	-0.5
Weighted mean 108.7							

^aFurther three-letter symbol: Asx, Glx = either acid or amide.

An amino acid residue, -HN-CHR-CO-, is the part of an amino acid (Figure 1-2a) occurring within a peptide chain; R denotes the side chain (Figure 1-1). Further one-letter symbol; B = Asx, Z = Glx, X = undetermined or nonstandard amino acid residue.

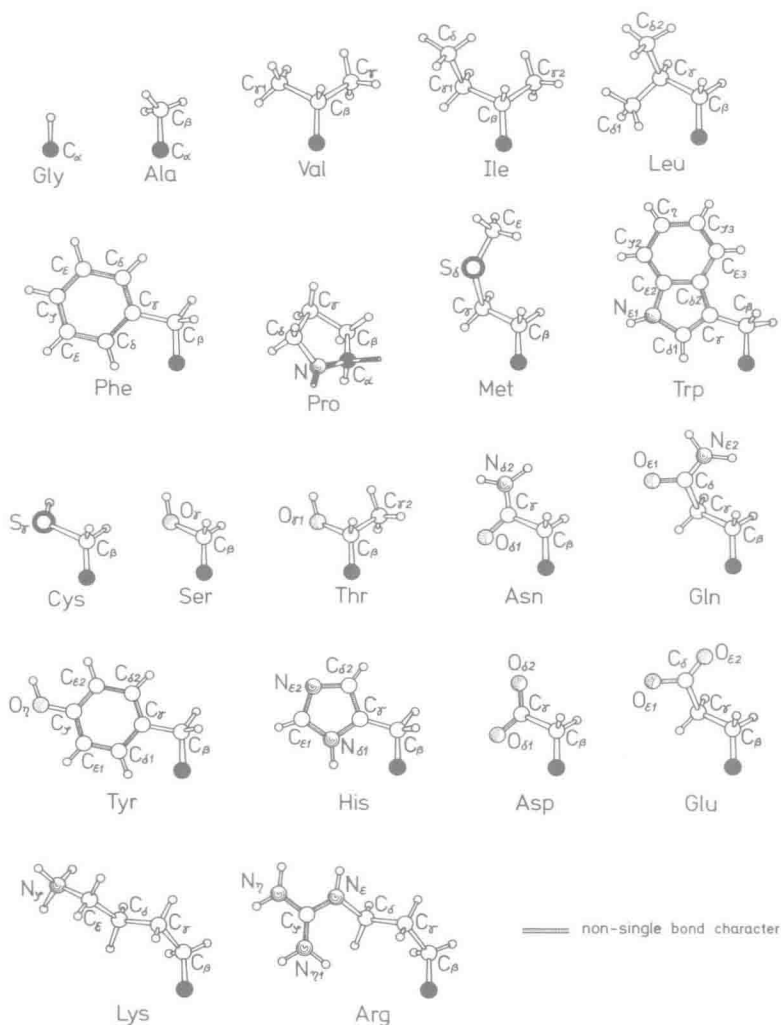


Figure 1-1. The 20 standard amino acid side chains. For proline part of the main chain is inserted. All other side chains are shown as they emerge from the C_α -atom in the main chain. The residue names are given as three-letter symbols. Atom names are those given in the IUPAC-IUB recommendations of 1969 (21). The main chain in Pro is indicated by solidly drawn bonds. All C_α -atoms are black.

started with single autocatalytic, that is, self-reproducing, molecules (3). Nucleic acids, and not proteins, are the candidates of choice for this role (4–6), because a single-stranded nucleic acid can use base pairing to align its complementary sequence of nucleotides. On linking the aligned nucleotides both strands can form the energetically favorable Watson–Crick double helix. Thus, linkage may be facilitated. In the next cycle, after

reopening the double helix, the complementary strand can reproduce the original strand. With proper sequence selection by base pairing and with the double helix structure providing a free energy incentive, autocatalytic reproduction is conceivable. In contrast, starting with proteins instead of nucleic acids would be much more difficult from an organizational point of view because no mutual recognition among amino acids exists.

Our knowledge of subsequent molecular evolution is also restricted to informed guesses. Although self-reproduction must have been very inefficient at the beginning, the molecules involved showed a tendency to improve autocatalysis, because they were subjected to strong evolutionary pressure; only the most efficient systems survived. Presumably, the improvement started with “enzymes” made of nucleic acid, splinters of the autocatalytic molecule itself. It is likely that ribosomal and transfer RNA are remnants of such catalysts. Later the nucleic acid enzymes were superseded by more efficient protein enzymes.

A linear array of standard elements with standard linkage is simple. The protein enzymes developed using a very straight-forward organizational scheme. The amino acid sequence in the polypeptide chain of proteins is a colinear and unique representation of the nucleotide sequence of the underlying nucleic acid. Three adjacent nucleotides code for one amino acid (Figure 1-5b). Accordingly, the polypeptides are similar to nucleic acids in that they are *linear* chain molecules with standard elements and one standard linkage. This allows for a simple and universal nucleic acid reading and polypeptide synthesizing machinery. The simplicity of linear systems is reflected, for example, in computer technology, where information storage and retrieval is usually handled via one-dimensional arrays put in standard format on linear data carriers such as magnetic tapes.

In nucleic acids as well as in proteins not only the linkages but also the atomic groups forming the backbone of the chain are uniform; in polypeptides all 20 amino acid residues are of the α -type and have the L-configuration at the C_α -atom (Figure 1-2a). All differences and therefore all information are restricted to the rather short side chains.

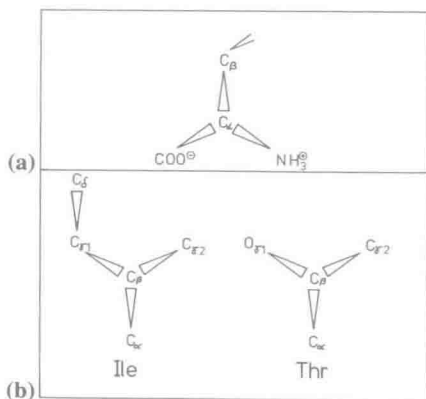


Figure 1-2. Asymmetric atoms occurring in standard amino acid residues. The view is from a hydrogen atom down to C_α or to C_β , respectively. This is opposite to the viewing direction in the Cahn-Ingold-Prelog presentation (41). (a) The standard L-isomer of an α -amino acid. The absolute configuration at C_α according to the Cahn-Ingold-Prelog notation is *S*. (b) The configuration at the C_β -atom of isoleucine (*S*) and at the C_β -atom of threonine (*R*).

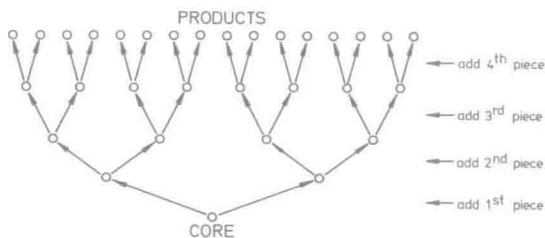


Figure 1-3. Hypothetical formation of molecular structures by adding various pieces to a core molecule, or by modifying a given precursor molecule specifically. In order to obtain different products each addition or modification must be specific, which means that it requires a specific enzymatic action. At the fourth stage $2^4 = 16$ different products are formed but $2 + 4 + 8 + 16 = 30 = 2^5 - 2$ manipulations were necessary.

Spontaneous folding is indispensable. After synthesis on the ribosome the polypeptide chain folds spontaneously to an amino acid sequence-dependent globular protein by adopting a state of lower free energy. It is possible that folding already starts during the synthesis. The resulting chain fold determines the specificity of the protein. This spontaneity simplifies the organizational scheme considerably, because otherwise “synthesizing enzymes” or “folding helper enzymes” would be required. Since very many chain folds are possible, a large number of such helping enzymes would be necessary.

Let us assume that structures were not built by spontaneous folding but, for example, by taking a small core molecule and linking subsequent pieces covalently to it until the desired form was attained. Even with the most efficient concept shown in Figure 1-3, such a process would require two enzymes for processing each intermediate and therefore an average of two enzymes per product molecule. Thus, in order to get a self-sustaining system all product molecules would have to be used as processing enzymes, each of which would have to catalyze two different synthetic reactions. However, in present day organisms generally only one enzymic action is observed for each molecule. More important, such a (super)tightly knit system would be very sensitive to errors—and it would have almost no chance to evolve from an unrefined primordial chemical soup (6). In conclusion, one appreciates the simplicity introduced by spontaneous folding.

But not all polypeptide chains fold spontaneously onto themselves. Some of the ribosomal proteins are very elongated and have very few internal contacts (7). Here, the chain assumes its final conformation by attachment to another molecule, namely the ribosomal nucleic acid core. These chains are probably very ancient and may reflect the situation which existed at the very beginning of protein production; ribosomes are highly conserved during evolution. This strong conservation is due to the ribosomes' central role in the life of an organism. Every change in the ribosome would be disruptive, giving rise to a very small chance for survival (3,4).

The standard amino acids can be produced under primordial conditions. The set of standard links in a polypeptide chain consists of 20 amino acid residues (Figure 1-1, Table 1-1). In order to be selected these amino acids must have been present in the environment of the primordial autocatalytic nucleic acid monsters. This was most likely the case, since several simulation experiments (3,8) have shown that amino acids can be produced under primordial conditions. In these experiments electric or radiation energy was pumped into an atmosphere considered representative of the atmosphere on early Earth. Usually, a racemic mixture of L- and D-amino acids (Figure 1-2a) of the α -type together with β -amino acids (Figure 1-4) and other compounds was obtained.

The L-configuration has been selected by chance. Is there any reason for the selection of the L-configuration at the C_α -atom? We do understand that all C_α -configurations are identical; the chain-synthesizing machinery requires a standard configuration, otherwise synthesis would be too complex and therefore would not have a competitive chance. But there is no obvious reason for preferring the L-configuration over its mirror image. Although it has been shown that the intrinsic asymmetry of β -decay expresses itself as molecular asymmetry by preferentially destroying D-amino acids (9), the observed effect of a few percent is too small to explain the selection. Presumably it occurred by chance and not because of a slightly higher amino acid concentration that the L-system developed first. Being farther advanced it then suppressed the emergence of a D-system. Alternatively, both systems developed in parallel, and the L-system won the ensuing competition by favorable environmental fluctuations, the impact of which presumably far exceeded the impact of a few percent difference in amino acid concentrations.

β -Amino acids are not suitable. Another aspect of these simulation experiments is the production of β -amino acids (Figure 1-4) in yields comparable to those of the α -type (8). In many respects β -amino acids could be used in the same way as the α -type except that the resulting polypeptide chain would contain C_α - C_β bonds which allow for additional free rotations. As shown in section 8.1 (page 151), such flexibility would prevent spontaneous chain folding. Thus, a reason for the selection of the α -type seems to be at hand.

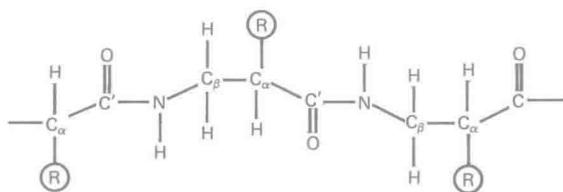


Figure 1-4. A hypothetical main chain consisting of β -amino acids. The rotation around the C_α - C_β and the C_β -N bonds would be rather free, rendering the chain very flexible. The argument would not change if the side chains R were attached to C_β instead of to C_α .