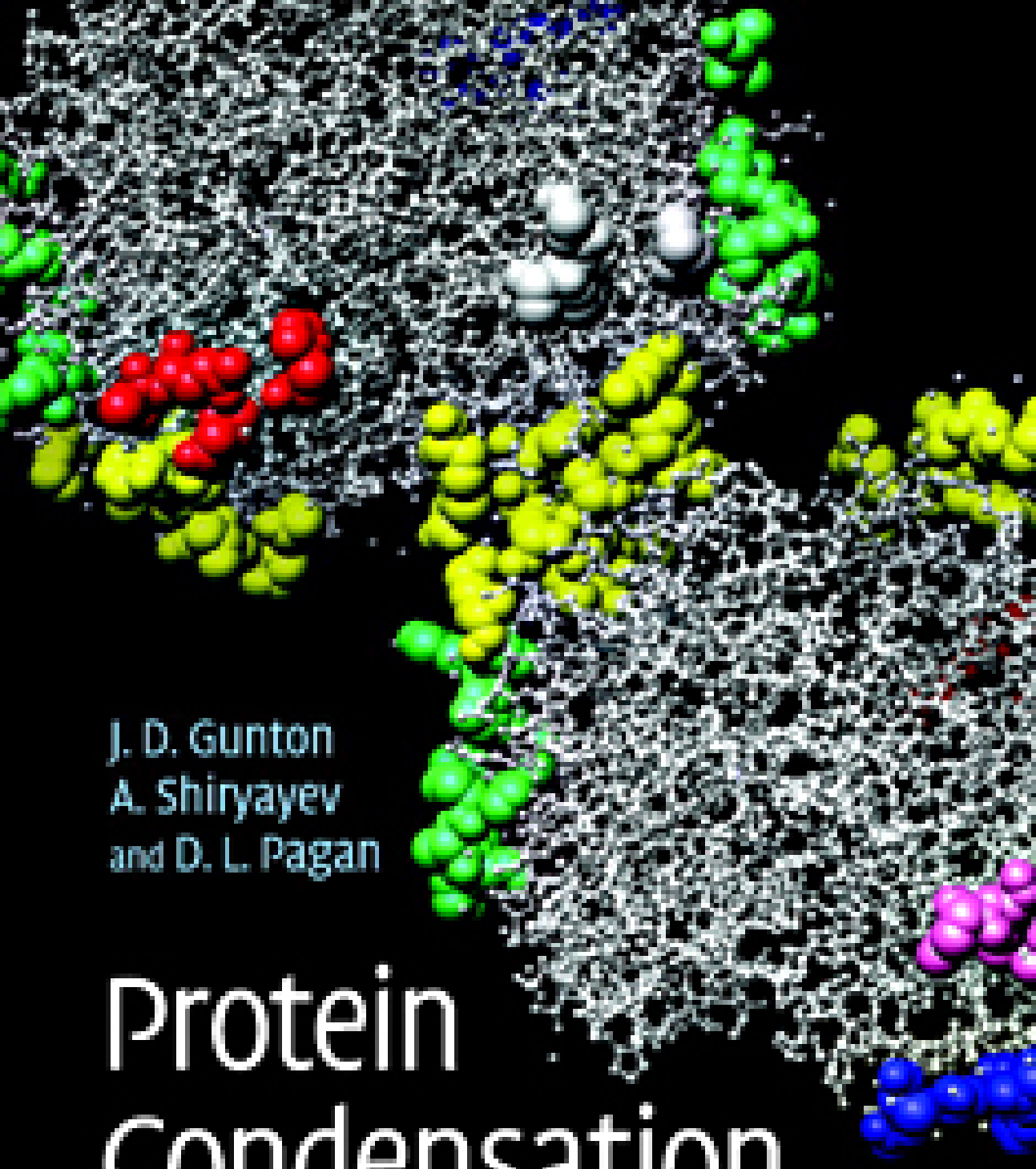


J. D. Gunton
A. Shiryayev
and D. L. Pagan

Protein Condensation

Kinetic Pathways to Crystallization and Disease

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521851213

This page intentionally left blank

PROTEIN CONDENSATION

Kinetic Pathways to Crystallization and Disease

This book deals with the phase transitions, self-assembly, and aggregation of proteins in solution. Its primary purpose is to bring an interdisciplinary audience the state of the art in current research. The book discusses issues related to the production of high quality protein crystals from solution, in which the bottleneck is crystal nucleation. Since protein function is determined by protein structure, it is necessary to grow high quality crystals in order to determine their structure, usually by X-ray crystallography. The main challenge is to determine the initial solution conditions so that optimal crystal nucleation occurs. The book also discusses diseases that occur due to undesired protein condensation, an increasingly important subject. Examples include sickle cell anemia, cataracts, and Alzheimer's disease. Current experimental and theoretical work on these diseases aims to understand the diseases at a fundamental, molecular level, in order to prevent the undesired condensation from occurring. Suitable for graduate students and academic researchers in physics, chemistry, structural biology, protein crystallography, and medicine.

J. D. GUNTON is Joseph A. Waldschmitt Professor of Physics at Lehigh University in Pennsylvania. He is the author of approximately 200 articles in refereed journals on equilibrium and nonequilibrium phase transitions. He is a Rhodes Scholar and a Danforth Fellow, and is a Fellow of the American Physical Society.

A. SHIRYAYEV and D. L. PAGAN received their Ph.D. degrees from Lehigh University in 2005. Both have already published several refereed articles that deal with the condensation of globular proteins.

PROTEIN CONDENSATION

Kinetic Pathways to Crystallization and Disease

J. D. GUNTON, A. SHIRYAYEV, AND D. L. PAGAN
Department of Physics
Lehigh University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK
Published in the United States of America by Cambridge University Press, New York
www.cambridge.org
Information on this title: www.cambridge.org/9780521851213

First published in print format 2007

ISBN-13 978-0-511-34007-9 eBook (Adobe Reader)
ISBN-10 0-511-34007-9 eBook (Adobe Reader)

ISBN-13 978-0-521-85121-3 hardback
ISBN-10 0-521-85121-1 hardback

To our wives, Peggy, Maria, and Martha
for their patience, love, and continued support

Contents

Preface

This book deals with a truly interdisciplinary subject: protein condensation from solution. We use "condensation" in this book to denote one of several forms of proteins: a dense, protein-rich fluid phase, an amorphous aggregate, a gel, a crystal, or a polymer fiber. All these forms have been observed experimentally and are important in their own right. The primary purpose of the book is to bring to a wide audience the current status of research in the field, which is still evolving at a rapid rate. The bulk of the book deals with issues related to producing high quality protein crystals from solution, in which the bottleneck is crystal nucleation. Here the main challenge is to determine the initial solution conditions so that optimal crystal nucleation occurs. A second and increasingly important subject that we discuss involves diseases that occur due to undesired protein nucleation. A classic example is the nucleation of polymer fibers of sickle hemoglobin molecules within the red blood cells that distorts the cells and produces sickle cell anemia. Another example is that of age-related cataracts produced by the undesired aggregation of γ-crystallin protein molecules within the vitreous fluid of the eye. A third, somewhat different, example involves the role of amyloid β protein in Alzheimer's disease. This list is likely to grow as scientists become more aware of the molecular origins of different diseases.

As the field is interdisciplinary, the first part of the book involves several brief reviews of subjects relevant to understanding protein condensation. Readers with an expertise in these topics should omit them and begin with the second part of the book, which treats several examples of globular and membrane proteins. The third part deals with the three diseases mentioned above.

We should note what the book does not deal with. It is not a treatise on the practical art/science of growing protein crystals. The classic work on this subject is A. McPherson's book, *Crystallization of Biological Macromolecules*. The emphasis in our book is on developing a statistical mechanics theory of the equilibrium and non-equilibrium aspects of protein condensation. We do not

discuss the kinetics of crystal growth since several review articles exist on this topic.

We would like to acknowledge the enormous help and encouragement we have received from our colleagues around the world. These include G. Benedek, F. Bonneté, F. Ferrone, T. Odijk, W. Poon, H. E. Stanley, A. Tardieu, D. Teplow, B. Urbanc, and P. Vekilov. We are particularly indebted to R. Sear for his critical reading of several chapters and his helpful, constructive comments. We also wish to thank I. Dokukina, M. Gunton, and N. Wentzel for their help with the manuscript.

Finally, one of us (JDG) would like to acknowledge funding from the Division of Materials Research of the National Science Foundation and the G. Harold and Leila Y. Mathers Charitable Foundation during the period we were writing this book. Without their support, this book would not have been written.

1

Introduction

1.1 Overview

This book deals with the condensation of proteins from solution, including protein crystal nucleation and certain diseases related to undesirable protein condensation.[1] We use the word condensation to describe a variety of possible states of matter, including dense, protein-rich fluids, amorphous aggregates, polymer fibers, gels, and crystals. Much of the book deals with understanding how to grow high quality protein crystals from aqueous solutions of protein molecules. This is of importance in structural biology, which deals with the study of the architecture and shape of biological macromolecules, and in particular with proteins and nucleic acids. Biologists are interested in knowing the structure of proteins, since structure determines function. To determine structure requires high quality protein crystals for use in X-ray crystallography. It is quite difficult to grow high quality protein crystals from solution, however; crystal nucleation is the major bottleneck in protein crystallography. Understanding the dependence of crystal nucleation on the initial conditions of the protein solution is a fundamental problem in statistical physics and is a major theme of this book. Understanding protein crystal nucleation is also important in biomedical research. For example, the sustained release of medications, such as insulin and interferon-α, depends on the slow dissolution rate of protein crystals [1–6]. One can obtain steady medication release rates for longer periods of time by using a dose of a few, larger, equidimensional crystallites than by a dose with a broad crystal size distribution. To obtain such a narrow size distribution requires an almost simultaneous nucleation of the crystals, so that the crystals can grow at the same decreasing supersaturation. In addition, as

[1] G. Benedek used "condensation" in the context of describing apparently unrelated diseases, including cataract, sickle cell anemia, and Alzheimer's disease [470]. He argued that these are all representations of a broad class of pathologies that he designated as "molecular condensation diseases"; these result from proteins condensing into dense, frequently insoluble phases. We extend this use of "condensation" to include states such as crystals, gels, and other aggregates.

discussed in subsequent chapters, certain diseases, such as sickle cell anemia and human cataracts, result from undesired protein condensation. In such cases one wishes to slow down or prevent the nucleation producing such condensation.

Protein condensation is an intellectually challenging subject in statistical mechanics, as there are many kinetic pathways for condensation to occur. Systems that are evolving toward their equilibrium states, which correspond to states with the lowest free energy, often get stuck in long-lived metastable intermediates. In many cases the outcome depends strongly on the initial position in the phase diagram. Several possible condensation states have been found to occur, such as those noted in the previous paragraph. In order to reach the desired outcome, one must understand the possible kinetic pathways; this is a formidable challenge. To obtain optimal crystallization, one must avoid gel and amorphous aggregate states and, instead, often take advantage of protein-rich liquid droplets via a metastable protein-poor, protein-rich liquid–liquid phase separation. To prevent sickle cell anemia from occurring in patients, one must prevent the nucleation of polymer fibers of sickle hemoglobin molecules from occurring while sickle hemoglobin is in its deoxygenated state in the cells. Although we are far from a detailed solution of the many problems discussed in this book, progress can be made by understanding the free energy landscapes for these systems. Indeed, it has been argued that one can use the free energy landscape of a system, normally used only for calculating its *equilibrium* properties, to predict the possible kinetic pathways that can occur in the course of a phase separation [7]. Although this does not in itself provide guidance on how to choose between various permitted pathways, it does at least limit the possibilities. Some progress has been made in obtaining the theoretical free energy landscape of a system by knowing its equilibrium phase diagram [7].

1.2 Protein function

Why study proteins? What makes them so important to human existence? The answer, of course, is that every living cell and all biological processes depend on proteins. This general statement reflects the fact that proteins are involved in every activity that is undergone in humans, or animals, on every level within the body. To illustrate their importance, consider the proteins involved in catalytic reactions, referred to by their special categorical name, *enzymes*. Examples of enzymes include pepsin, chymotrypsin, and trypsin, which are involved in the digestive process. Like all catalytic agents, enzymes accelerate the rates of chemical reactions while remaining intact themselves. The three aforementioned enzymes are produced in the mucosal lining of the stomach, and act to break down dietary proteins. These enzymes work together to simplify ingested proteins

into their fundamental components, which can then be easily absorbed by the intestinal lining. Interestingly, these digestive enzymes were among the first to be successfully crystallized, confirming an earlier finding that enzymes were indeed proteins. Another example of protein function is the way we protect ourselves from injury or disease. When we are wounded, the protein thrombin, along with other proteins and platelets, is activated in an attempt to form a clot, thereby preventing the loss of blood. Deficiency of these clotting factors is the cause of bleeding disorders such as hemophelia. When we are sick from bacterial or viral infection, our immune system responds by activating antibodies (the proteins of which are referred to as the immunoglobins) to fight off the invasion. Proteins are also involved in supporting the structure of our cells. Examples include collagen, found in tendons and cartilage, and keratin, which is found in hair and fingernails. As another example of their diversity, consider that proteins are needed in order to transport material. A prime example is hemoglobin, which is found in red blood cells and is responsible for transporting oxygen to living cells. Other obvious examples are nutrient proteins such as ovalbumin and casein, required by our bodies for proper growth and development. Thus, a whole consortium of proteins, with various functions and degrees of importance, exist in the body.

The extraordinary diversity of protein function is due to the precise specificity of a given protein's interactions with molecules. Molecules have to "fit" into the protein, which requires a relatively rigid spatial structure of the protein. The structure of a protein determines its function through determining the molecules with which the protein interacts. As a consequence, obtaining protein structure is a high scientific priority. Currently, X-ray crystallography is the primary method to determine structure; this requires high quality protein crystals. The growth of such crystals from supersaturated solutions of protein in solvent depends sensitively on the initial conditions of the solution. Until relatively recently, finding the particular initial conditions requisite for optimal crystal nucleation from solution was a trial and error process. Considerable progress has been made in understanding the role of the initial conditions, however. It turns out that *metastable* fluid–fluid critical points play a key role in determining optimal crystal nucleation, as we discuss in Chapter 7 and elsewhere.

The condensation of globular proteins is also a crucial factor in certain human diseases. The completion of the human genome project has brought with it a huge inventory of information regarding identification of genes, and, in addition, has helped to usher a change in philosophy regarding our outlook on disease and its treatment. Scientists are increasingly considering disease at a molecular level in order to understand the causes and aid in the prevention of certain diseases. Given the abundance and diversity of proteins, it is not unreasonable to view them simultaneously as the cause and treatment of disease. Scientists have begun using

proteins to test how a person's system reacts when foreign proteins are injected into the body. This could be caused by genetic factors and may be involved in the development of such diseases as diabetes mellitus and hypertension. Recent studies on proteins in the body have also shed light on the causes of some diseases. It has been shown (see Chapter 13) that a genetic mutation in the hemoglobin molecule, HbS, is involved in sickle cell anemia. One study linked a liquid–liquid phase transition with the polymerization of the molecule, the precursor of the disease which ultimately gives the red blood cells their signature sickle shape. Other studies (see Chapter 12) have shown that genetic cataracts are also caused by protein crystallization in the eye lens, effectively clouding the transparency of the eye. Alzheimer's disease has also been linked to the crystallization of the protein molecule amyloid β protein (see Chapter 14).

Other uses of proteins stem from applications in the pharmaceutical industry. Protein crystals are being studied for use in vaccine delivery [8]. Drugs are also being designed to attach to protein sites of infected cells to deliver drugs. In another application of drug delivery, proteins themselves are being used to help combat disease such as hepatitis C. The protein molecule interferon, along with the help of a process dubbed pegylation, targets infected cells and delivers medicine to treat the patient. Protein crystals are also becoming useful in biotechnology. For example, the stabilization of enzymes as industrial catalysts involves crystallization of the enzyme followed by a subsequent cross-linking [9].

1.3 Types of proteins

The two major classes of proteins are the fibrous proteins and globular proteins. Fibrous proteins are abundant in cells and perform tasks that require each protein molecule to span a large distance. These have a relatively simple, elongated structure. We will not concern ourselves with this class of proteins in this book. Globular proteins are compact and approximately spherical in shape, with an irregular surface. They are by far the most numerous of cellular proteins and, unlike fibrous proteins, tend to be soluble in aqueous media. Globular proteins comprise most of the structures in the protein data bank. These proteins perform most of the chemical functions of the cell, including synthesis, transport, and body metabolism. Examples of this class include all the enzymes, albumin, globulin, casein, hemoglobin, and protein hormones. Hemoglobin is a respiratory protein contained in red blood cells and carries oxygen throughout the body. There are more than 100 different forms of human hemoglobin, including hemoglobin S, the cause of sickle cell anemia, which is the subject of Chapter 13.

Membrane proteins form a third class of proteins. These are proteins that are associated with the lipid bilayer of a cell membrane and carry out most of the

membrane functions. Many membrane proteins extend through the bilayer, with both hydrophobic and hydrophilic regions. Their hydrophobic regions lie in the interior of the bilayer, in contact with the hydrophobic tails of the lipid molecules. Their hydrophilic regions are exposed to the water environment on either side of the membrane. Other membrane proteins are located completely outside the bilayer, being attached to the bilayer only by one or more covalently attached lipid groups. Finally, others are attached to the membrane only relatively weakly, via non-covalent interactions with other membrane proteins. Chapter 11 deals with membrane proteins.

1.4 Protein crystallization

Crystallization is usually induced in the laboratory by adding salt, alcohol or polymer to dilute protein solutions [10]. These methods were developed from the first studies of crystallization, which took place over 150 years ago. Surprisingly, the pioneering methods such as dialysis of salt solution and the use of organic solutions as precipitating agents are still used today and comprise the basic tools of crystal growth. These early successes showed that protein crystals behave in much the same way as inorganic crystals and were used to provide "proof" of the purity of the sample from which the crystals were obtained. These also demonstrated that crystals could be obtained from solution. Figure 1.1 shows hemoglobin crystals, first successfully crystallized in 1840. With the advent of X-ray diffraction, and subsequently other techniques such as light scattering and NMR spectroscopy, protein crystals are providing the means by which scientists can determine the structure of proteins. Even with these techniques, however, good crystals that are free from defects are not easily produced. The basic conditions under which good quality (suitable for X-ray crystallography, for example) crystals can be obtained are not understood. Methods of growing crystals currently depend on trial and error tests to see under what conditions a protein solution will crystallize. Also, because each protein is different, conditions for which one particular solution will result in protein crystals do not result in crystals for another protein solution. To overcome these problems, crystallographers implement a "factorial" method. This is a brute force means in which all possible variants of initial conditions are examined. Another promising method that is currently being pursued involves the use of microfluid techniques.

Of the many factors that govern protein crystallization, it is known that a super-saturated solution promotes crystal growth. The most popular of the precipitates available to promote supersaturation is poly-ethanol glycol (PEG), a long flexible polymer chain. This enhances precipitation due to a depletion effect, discussed in Chapter 5. PEG can be grown with various lengths. Crystals grown out of

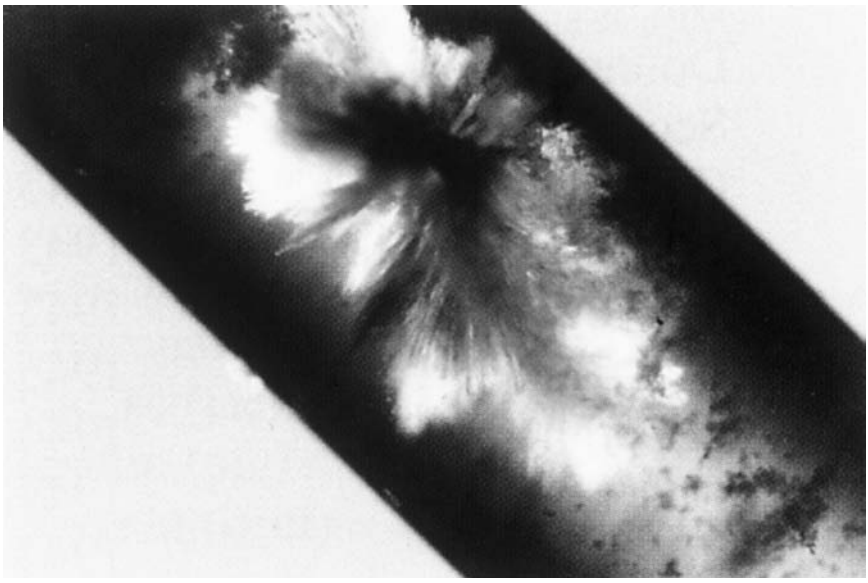


Figure 1.1. Fine needle crystals of hemoglobin, much like those grown by Hünfeld in 1840. Reprinted, with permission, from ref. [10]. For a color image, please see Plate Section.

solutions that contain PEG have been shown by X-ray structure analysis to be the same as those grown by traditional methods. An advantage of PEG over other precipitants is that most macromolecules crystallize within a fairly narrow range of PEG concentration. Recently, many studies have examined the influence of PEG on the phase diagrams of protein solution, as we discuss in later chapters.

1.5 Outline of book

This book is divided (roughly) into three parts. Chapters 2 to 5 review several topics relevant to protein discussion, including protein structure, experimental methods, thermodynamics and statistical mechanics, and protein–protein interactions. The subject of the interactions between protein molecules in solution is fundamental to our ability to calculate phase diagrams and non-equilibrium properties such as nucleation rates. Our theoretical understanding of these interactions, however, is still incomplete, so our discussion is incomplete. There is no doubt that further progress in understanding the role of the solvent (typically water, buffer, and precipitants such as salts or PEG) in determining the interactions between the protein molecules is crucial to our ability to understand and control the relevant equilibrium and kinetic properties of these protein solutions. This is clearly a major subject for future research.

Chapters 6 to 11 deal with a variety of topics that have been the subject of extensive research. Chapter 6 summarizes results for the microscopic models that have been used to model protein solutions and discusses various simulation and theoretical tools that are available for these studies. Chapters 7 and 8 review our current theoretical and experimental studies of nucleation. Both topics require further development; in particular, there are very few quantitative measurements of homogeneous crystal nucleation rates. Since nucleation is the bottleneck in protein crystallization and is also of fundamental importance in several diseases, this is another major subject for future research. Chapters 9, 10, and 11 discuss our experimental and theoretical understanding of lysozyme, some additional globular proteins, and membrane proteins, respectively. Lysozyme is by far the best studied globular protein; it is fair to say that its equilibrium properties are now well characterized experimentally as a function of several control parameters such as salt type and concentration of salt and PEG. Progress has also been made in understanding its non-equilibrium properties, including its diffusion constant and nucleation rate, although there is room for further experimental work here. But lysozyme is by no means typical of globular proteins; therefore, Chapter 10 summarizes results for several other proteins, including urate oxidase, alpha crystallin, ATCase and apoferritin. We also summarize an impressive study of protein crystallization involving liquid–liquid phase separation for glucose isomerase.

Chapters 12 to 14 treat three examples of disease that involve protein condensation. Chapter 12 discusses the relationship between crystallins and a class of age-related cataracts, as well as current theoretical efforts to model this. Chapter 13 summarizes our experimental and theoretical understanding of sickle hemoglobin and its role in sickle cell anemia. This chapter deals with a different type of nucleation process than those for other proteins. In this case, a complex polymerization of fiber chains from sickle hemoglobin molecules plays a crucial role. This process is thought to occur via a two-step mechanism of homogeneous and heterogeneous nucleation. Finally, Chapter 14 reviews the state of experimental and theoretical understanding of the role of amyloid β protein in Alzheimer's disease. Each of these chapters is an enormous area of research, so our discussion is, by necessity, incomplete. We provide references of several major reviews of these topics for further study.

2

Globular protein structure

All proteins are linear polymers of amino acids, large sequences of which constitute a peptide chain. Our focus is on globular proteins, whose peptide chain has a folded structure. In general, they are soluble in water and in other polar solvents. Although their structure is complex, we will see in this chapter that they often assume similar forms and that their shapes, sequence, and conformation can be understood by considering some fundamental aspects of their structure.

2.1 Amino acids and primary structure

The fundamental unit (monomer) of the protein molecule is the α amino acid. It consists of an acidic carboxyl group and an amino group attached to a single carbon atom, referred to as the α-carbon, and a hydrogen atom. This is illustrated in Fig. 2.1. A side-chain of molecules, designated as "R," is also attached to the amino acid. This side-chain is specific to each amino acid and is what differentiates them from each other. The side-chains can vary in complexity; examples are simple hydrogen atoms, an extra amino group, an extra carboxylic group, a sulphydryl group, a hydroxyl group, or a simple hydro-chain or hydro-carbon ring. There are 20 biologically important amino acids listed in Table 2.1. The table also lists some proteins and their amino acid composition.

In order to form the larger protein molecules, these amino acids need to be "linked" together. This is accomplished by means of a *peptide* bond. This bond occurs between the carbon atom of the carboxyl group in one amino acid and a nitrogen atom in the amino group of another. Figure 2.2 shows a peptide bond between two amino acids, forming a dipeptide; three amino acids would constitute a tripeptide. Protein molecules consist of many amino acids bonded together, and are appropriately referred to as polypeptides. The varieties that can occur in a single polypeptide are enormous. Recall that there are 20 biologically useful amino acids. Already in forming a dipeptide there are 400 possible such

Table 2.1. *Amino acid composition of some selected proteins*

Values expressed are percent representation of each amino acid. RNase: bovine ribonuclease A, an enzyme. ADH: horse liver alcohol dehydrogenase; the amino acid composition of this protein is reasonably representative of the norm for water-soluble proteins. Mb: sperm whale myoglobin, an oxygen-binding protein. Histone H3: histones are DNA-binding proteins found in chromosomes. Collagen: collagen is an extracellular structural protein.

Amino acid	RNase	ADH	Mb	Histone H3	Collagen
Ala	6.9	7.5	9.8	13.3	11.7
Arg	3.7	3.2	1.7	13.3	4.9
Asn	7.6	2.1	2.0	0.7	1.0
Asp	4.1	4.5	5.0	3.0	3.0
Cys	6.7	3.7	0	1.5	0
Gln	6.5	2.1	3.5	5.9	2.6
Glu	4.2	5.6	8.7	5.2	4.5
Gly	3.7	10.2	9.0	5.2	32.7
His	3.7	1.9	7.0	1.5	0.3
Ile	3.1	6.4	5.1	5.2	0.8
Leu	1.7	6.7	11.6	8.9	2.1
Lys	7.7	8.0	13.0	9.6	3.6
Met	3.7	2.4	1.5	1.5	0.7
Phe	2.4	4.8	4.6	3.0	1.2
Pro	4.5	5.3	2.5	4.4	22.5
Ser	12.2	7.0	3.9	3.7	3.8
Thr	6.7	6.4	3.5	7.4	1.5
Trp	0	0.5	1.3	0	0
Tyr	4.0	1.1	1.3	2.2	0.5
Val	7.1	10.4	4.8	4.4	1.7
Acidic	8.4	10.2	13.7	8.1	7.5
Basic	15.0	13.1	21.8	24.4	8.8
Aromatic	6.4	6.6	7.2	5.2	1.7
Hydrophobic	18.0	30.7	27.6	23.0	6.5

From ref. [11]. Reprinted with permission of Brooks/Cole, a division of Thomson Learning: www.thomsonrights.com.

peptides. Polypeptides then can be vastly different in regard to their amino-chain composition. This gives rise to the different proteins that are found in nature and underlies their diversity and abundance.

The sequence of amino acids in a protein molecule determines the peptide backbone or *primary structure*. It is encoded by the nucleotide sequence in DNA and constitutes a form of genetic information. Note that the primary structure only refers to the sequence of its monomers without any regard to the side-chains. Ultimately, the behavior of these molecules is determined by the shape they conform to in three-dimensional space.

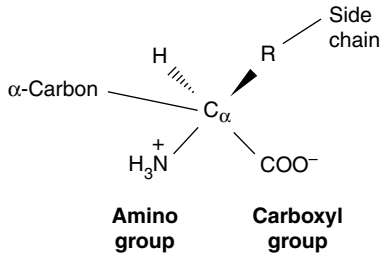


Figure 2.1. Anatomy of an amino acid. Except for proline and its derivatives, all of the amino acids commonly found in proteins possess this type of structure. From ref. [11]. Reprinted with permission of Brooks/Cole, a division of Thomson Learning: www.thomsonrights.com.

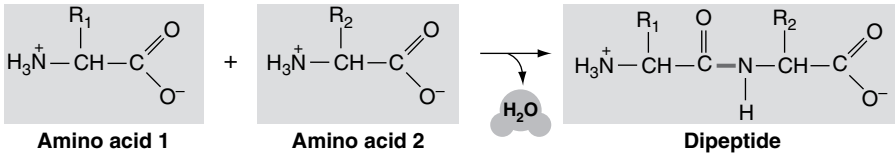


Figure 2.2. Peptide formation is the creation of an amide bond between the carboxyl group of one amino acid and the amino group of another amino acid; R_1 and R_2 represent the R groups of two different amino acids. From ref. [11]. Reprinted with permission of Brooks/Cole, a division of Thomson Learning: www.thomsonrights.com.

2.2 Secondary structure

The conformation of the peptide backbone is referred to as the secondary structure and is dominated by hydrogen bonding. This is attributable to the partial negative charges on the oxygen and nitrogen atoms, and the positive charge on the hydrogen atoms. As a result, the peptide atoms arrange themselves so as to accommodate these attractions. However, this arrangement is limited by steric considerations and, more importantly, by the restrictions placed upon them by the peptide bond itself. Generally, the bond between the oxygen and carbon atoms is drawn as a double bond and the peptide bond is drawn as a single bond. This, however, is not accurate. In reality, the electrons on the nitrogen and oxygen atoms are delocalized, and are "shared" among the three atoms. This has some important consequences. One result is that the peptide bond is longer than a single bond, but shorter than a double bond, having a length of 0.133 nm. A more important consequence is that the delocalization restricts the rotation of the bonds. Specifically, the six atoms involved in the peptide link are forced into a planar structure. Figure 2.3 shows the six atoms arranged in this planar structure. The α-carbon bonds with the nitrogen and carbon atoms are not restricted in this way, but can be restricted by steric limits. The peptide links can be viewed as planar sheets which can be

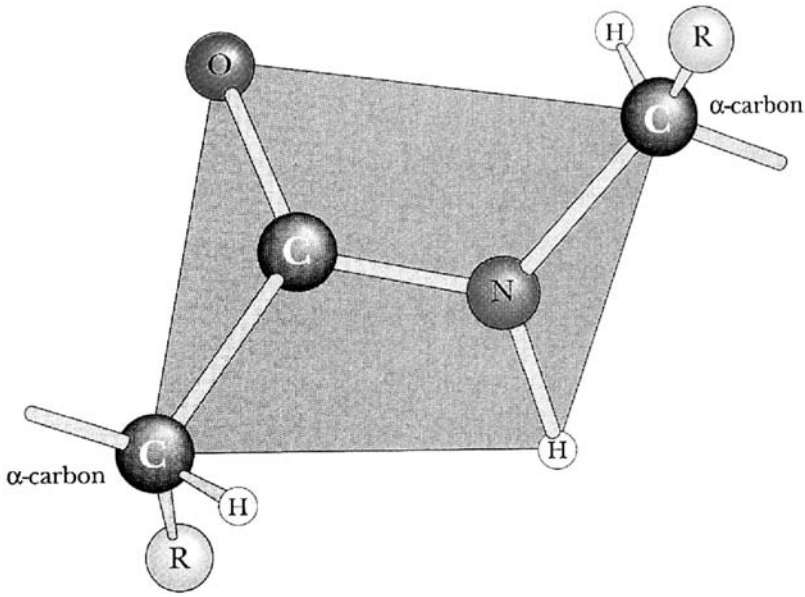


Figure 2.3. The coplanar relationship of the atoms in the amide group is highlighted as an imaginary shaded plane lying between two successive α-carbon atoms in the peptide backbone. From ref. [11]. Reprinted with permission of Brooks/Cole, a division of Thomson Learning: www.thomsonrights.com. For a color image, please see Plate Section.

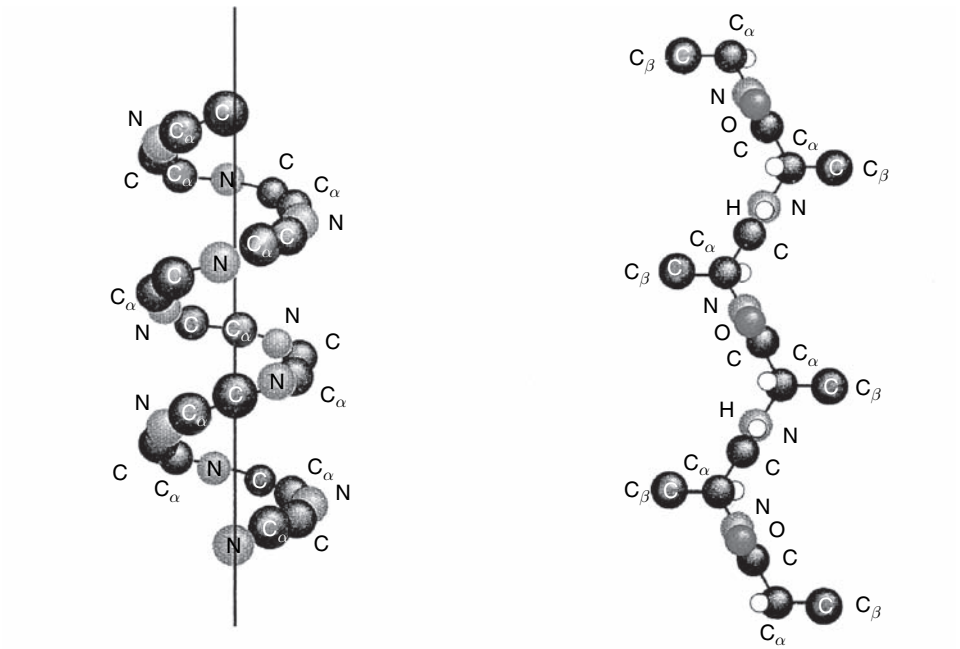
manipulated to accommodate hydrogen bonding. There are many different ways in which these sheets can align themselves. When these sheets have identical rotations or orientations, they give rise to particular secondary structures. Two common structures are the α-helix and β-sheet.

As its name suggests, the atoms in the α-helix arrange themselves in a helical pattern. Figure 2.4 shows a helical arrangement of the atoms in this structure. Every nitrogen atom avails itself to hydrogen bonding. In fact, if we consider the hydrogen atoms themselves, we see that every amide hydrogen and carbonyl oxygen is involved in a hydrogen bond. That every peptide link is involved in two hydrogen bonds makes this structure very stable. The linear arrangement of the hydrogen-bonding atoms is such that the bonds are nearly at their maximum strength.

Instead of the peptide links arranging themselves via twists and turns, they can bond with other strands of the polypeptide in a sheetlike fashion. Also shown in Fig. 2.4 are two segments of a polypeptide chain linked side by side in a parallel fashion (with their amino groups and carboxylic groups at the same ends of the chains). In this conformation, referred to as the beta sheet (or β-sheet), the nitrogen–hydrogen and carbon–oxygen bonds point out at right angles to the peptide backbone, each set of atoms alternating alongside it.

α-Helix
Only the N—C_α—C backbone is represented. The vertical line is the helix axis

β-Strand
The N—C_α—C_O backbone as well as the C_β of R groups are represented here. Note that the amide planes are perpendicular to the page.



"Shorthand" α-helix

"Shorthand" β-strand

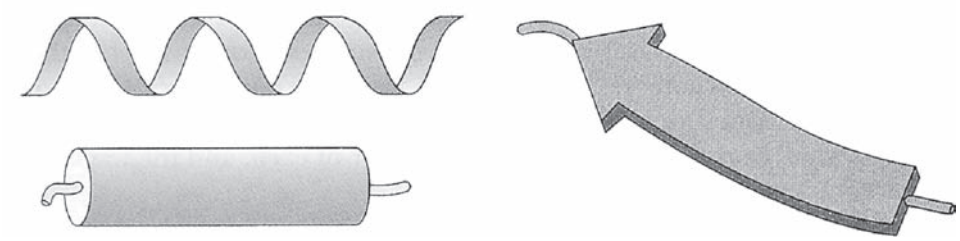


Figure 2.4. Two structural motifs that arrange the primary structure of proteins into a higher level of organization predominate in proteins: the α-helix and the β-pleated strand. Atomic representations of these secondary structures are shown here, along with the symbols used by structural chemists to represent them: the flat, helical ribbon for the α-helix and the flat, wide arrow for β-structures. Both of these structures owe their stability to the formation of hydrogen bonds between N–H and O=C functions along the polypeptide backbone. From ref. [11]. Reprinted with permission of Brooks/Cole, a division of Thomson Learning: www.thomsonrights.com. For a color image, please see Plate Section.

To distinguish the different classes of proteins, namely globular proteins and fibrous proteins, the secondary structure is generally different. In fibrous proteins, their polypeptide chains have extensive regions of regular secondary structure, consisting mainly of bundles or sheets. This is not surprising since these proteins are usually involved in maintaining the strength or structure of cells. Globular proteins, in contrast, do not have such large regions of secondary structure. Their backbones are arranged with much less regularity.

2.3 Tertiary structure

In addition to the conformation of the peptide backbone, the side-chains too must conform in three-dimensional space. Typical examples are shown in Figs. 2.5 and 2.6. These interactions and the interactions between the secondary structures are largely responsible for the folding of proteins. These interactions are governed by the hydrophobic/hydrophilic effects, hydrogen bonding, electrostatic, and van der Waals interactions.

Figure 2.5. Tertiary structure of the protein molecule serum albumin. From ref. [12]. For a color image, please see Plate Section.

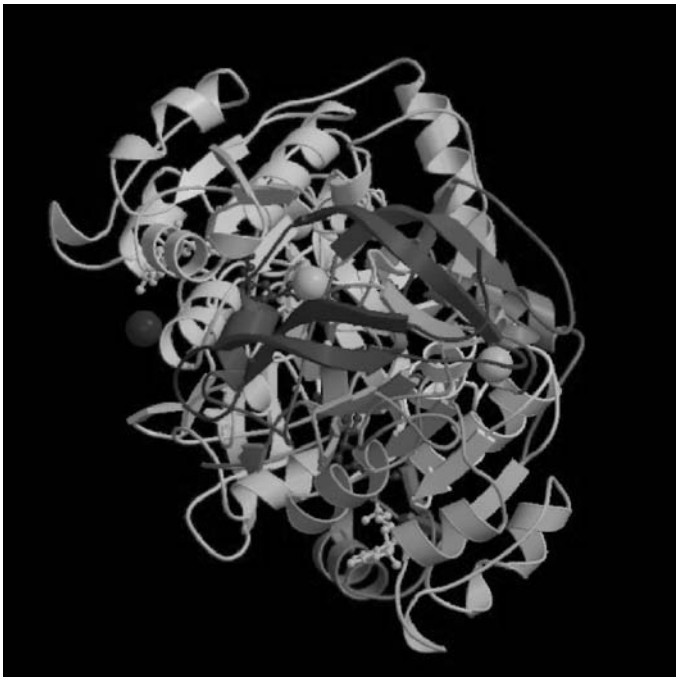


Figure 2.6. Tertiary structure of the protein molecule alcohol dehydrogenase. From ref. [12]. For a color image, please see Plate Section.

The hydrophobic effect plays a dominant role in the folding of proteins. This effect describes the interactions between non-polar, neutral molecules and their inability to react with water or any other polar solvent. Imagine water as a (loose) network of polar molecules. Over time, these molecules attempt to orient themselves to account for the attractions between positive and negative regions of the molecules. If a non-polar molecule, or a group of them, enters this system, there is no orientation of the molecule that will allow it to bond to the water molecules. As a net result, these molecules are pushed together and are excluded from the "network" altogether, forming their own region. On the opposite spectrum, the hydrophilic effect corresponds to interactions between particles that have an affinity to react with water or any polar solvent. Polar molecules, such as Na^+ or Cl^-, readily react and mix well with water. Thus, the water bonds are no longer disrupted, since these hydrophilic molecules can easily mix in with the solvent. With regard to proteins, polypeptides can contain numerous amino acids. These molecules can contain hydrophobic or hydrophilic regions, or both. Generally, it is common for protein molecules to consist of monomers of both groups. As a result, in an aqueous environment, the molecules orient themselves such that the

hydrophilic regions are exposed on the surface, where they can react with the solvent. The interior of the molecule is composed of a hydrophobic core.

Hydrogen bonding, which is responsible for the secondary structure of protein molecules, can occur between the side-chains as well. Polypeptide chains have numerous (hydrogen) donors and (oxygen) receptors. They can also react with the environment. Electrostatic interactions occur in three main types: permanent dipole–permanent dipole, permanent dipole–induced dipole, and induced dipole–induced dipole interactions. van der Waals forces involve reactions between induced dipoles with each other, as discussed in Chapter 5 on protein interactions.

Protein function is largely determined by its tertiary structure. Antibodies, for example, can bond to antigens because of the special binding sites adopted by them. If an antibody were to lose its structure, through genetic mutation perhaps, it would lose this ability. This loss of protein structure is referred to as *denaturization*, the state in which it has lost its ability to function properly, and can occur by other means as well. Indeed, experimentalists must take proper measures to ensure that the structure of protein molecules is not destroyed in an experiment.

Figure 2.7. Quaternary structure of the protein molecule lectin agglutinin. From ref. [12]. For a color image, please see Plate Section.

2.4 Quaternary structure

Many protein molecules consist of several (usually identical) polypeptide chains bound and arranged to form a larger macromolecule. The peptide chains constitute a subunit of the molecule, and protein molecules can have a different number of subunits. For example, aggregates of two subunits are known as dimers, those of three subunits are trimers, etc. This aggregation of subunits form the quaternary structure of the protein molecule. Figure 2.7 shows an example of the quaternary structure of a protein molecule.

The subunits of a protein molecule are not randomly arranged but are related by symmetry relations. For example, positions of the subunits of a dimer are related by a 180° rotation about an axis perpendicular to the plane of the paper, known as a two-fold axis. In general, a rotation of $360/z$ defines a z-fold axis. Not all subunits of a protein need to be identical to form a quaternary structure. The well studied protein hemoglobin is a tetramer, having four subunits, with two kinds of identical peptide strands labelled as α_j and β_j, with $j = 1, 2$. There is a two-fold symmetry among the α- and β-strands.

3

Experimental methods

In this chapter we provide a brief summary of some of the methods used to study the physical properties of proteins in solution. This is not meant as a thorough discussion of experimental techniques, but rather as an introduction with sufficient references to provide the interested reader with a guide to the literature.

3.1 Methods to determine three-dimensional protein structure

As noted earlier, knowing the structure of a protein molecule is crucial to understanding its biological function. Over the past century, efforts have been focused on methods which allow protein structure to be determined. X-ray crystallography and NMR studies are two popular techniques, the former being predominant.

3.1.1 X-ray crystallography

X-ray crystallography [13, 14] of protein crystals is a well known technique, which we briefly summarize here. Its use for protein crystals, of course, depends upon the regular arrangement of protein molecules on a crystal lattice. When electromagnetic radiation of a given wavelength, such as X-rays, is incident upon a grating of appropriately sized spacings, the waves interfere constructively and destructively in a process known as diffraction. The diffraction pattern is a three-dimensional image of the crystal in reciprocal space and can be used ultimately to determine the real structure of the lattice; it represents the scattering of the X-rays by the electrons of the atoms in the crystal. From the structure factor one can readily obtain the electron density map of the crystal structure. The structure factor is given by

$$\mathbf{F(S)} = \int \rho(\mathbf{r}) \exp(\mathrm{i}\phi)\, \mathrm{d}V, \tag{3.1}$$

where $\rho(\mathbf{r})$ is the electron density of the crystal, $\phi = 2\pi\mathbf{r}\cdot\mathbf{S}$ is the phase difference or phase angle (with $\vec{S}$ the resultant scattered wave) of the scattered radiation, and the integration is over all space. However, scattering only occurs if all of the unit cells scatter in phase. This places restrictions on the values that the phase angle $\phi = 2\pi\mathbf{r}\cdot\mathbf{S}$ can take in accord with the Laue condition. Also, determining the phase angle is difficult. A more useful form of the structure factor is given by

$$\mathbf{F(S)} = \int \rho(\mathbf{r}) \exp(2\pi \mathrm{i}[hx + ky + lz])\mathrm{d}V, \tag{3.2}$$

where the phase angle has been written in terms of Miller's indices – the integers (h, k, l) – and the atomic coordinates of the crystal (x, y, z). The electron density is important to crystallographers and can be obtained via a Fourier transform of the structure factor, yielding

$$\rho(\mathbf{r}) = \frac{1}{V}\sum_h \sum_k \sum_l \mathbf{F}(h, k, l) \exp(-2\pi \mathrm{i}[hx + ky + lz]). \tag{3.3}$$

The summation reflects the fact that the structure factor is not continuous and is non-zero only at the reciprocal lattice points (h, k, l) as described by the Laue conditions. A major obstacle in protein crystallization (and in crystallography, in general) is obtaining the phase of the structure factor. Noting that $F(h, k, l)$ is a complex number, Eq. (3.3) can be rewritten as follows:

$$\rho(x, y, z) = \sum |F| \exp(-2\pi \mathrm{i}[hx + ky + lz] + \mathrm{I}\alpha(h, k, l)), \tag{3.4}$$

where α is the phase of the structure factor. Though knowledge about the magnitude of the structure factor is gained, all information regarding its phase is lost. This is known as the *phase problem* in crystallography. The most widely used method of obtaining the phase angle is by use of the isomorphous replacement method. The isomorphous replacement method requires the diffraction pattern of the native protein crystal, and that of the same crystal with the addition of heavy atoms (high atomic number) introduced into the protein molecules. The idea is to obtain two identical diffraction patterns, with that of the derivative having a higher intensity. For perfect isomorphism, the structure and conformation of the native protein crystal and the heavy-atom derivative must be the same. The changes in the intensity of the diffraction pattern are then wholly due to the presence of the inserted heavy atoms. Determination of the positions of the heavy atoms allows for the determination of the phase angles.

3.1.2 NMR spectroscopy

NMR spectroscopy offers another means by which protein structure can be elucidated. This technique involves proteins in solution, and has primarily been

used to determine the structure of small proteins, involving 200 or fewer amino acids. The use of NMR often requires adding an isotope such as C^{15} or N^{15} with fewer than 200 amino acids. NMR has also been used to measure the induction time for lysozyme crystallization at different temperatures [15]. The induction time is the time interval between the moment supersaturation has been established and the later moment at which crystal nuclei first appear.

3.2 Solubility measurements

When a homogeneous solid, such as a protein crystal, is added to a solvent such as water, it will in general begin to dissolve. This dissolution process will continue until the rate at which molecules entering the solution from the solid phase (solute) equals the rate at which the molecules leave the solution to become solid. The concentration of the substance in solution at this point is its equilibrium, or solubility, value. The locus of these concentration values as a function of temperature (holding the other solvent parameters constant) is known as the solubility curve (or solubility line or liquidus line) of that material. If the concentration of the material is less than its equilibrium value, the solution is said to be undersaturated; if it is greater than its equilibrium value, the solution is said to be supersaturated.

For many systems the equilibrium solubility concentration increases with increasing temperature; this is a normal solubility curve. For some systems the solubility concentration decreases with increasing temperature; this is a retrograde solubility curve. In a few cases the solubility concentration is essentially independent of temperature and appears as a vertical line (approximately) in the temperature–concentration phase diagram.

There are various ways to determine the solubility curves for proteins. A standard method is to place the protein crystal in a protein-free solution at some specified conditions of temperature, pH, etc. As the crystal dissolves, the solution is stirred to mix its components. The concentration of the solution is measured by removing a small sample and determining its concentration by some means, such as UV absorption spectroscopy. At some point in the process, the crystal will stop dissolving, as long as the volume of the solvent is sufficiently small. The solute concentration at this point is the solubility limit; the solubility curves are known for more than 40 proteins [16]. A rapid determination of protein solubility diagrams has been developed that employs a miniature column technique [17].

3.3 Second virial coefficient

We digress in this discussion of experimental techniques to define a quantity, the osmotic second virial coefficient, that is a measure of the interaction between

proteins in solution. The osmotic pressure Π of solute particles in solution can be expanded in terms of the protein number density ρ as follows:

$$\frac{\Pi}{\rho kT} = 1 + B_2\rho + \cdots, \tag{3.5}$$

where k is the Boltzmann constant, T is the absolute temperature, and B_2 is the second virial coefficient. Alternatively, one can expand the osmotic pressure in terms of the concentration to obtain an equivalent form:

$$\frac{\Pi}{cRT} = \frac{1}{M} + A_2 c + \cdots, \tag{3.6}$$

where c is the protein concentration ($\mathrm{g\,cm^{-3}}$), M is the protein molecular weight ($\mathrm{g\,mol^{-1}}$), and $\rho = cN_\mathrm{A}/M$, where N_A is Avogadro's number; A_2 is also known as the second virial coefficient (with different units than B_2, of course), so that one must be careful to distinguish which second virial coefficient one is referring to in a particular experiment. (Note: to make matters even more confusing, some authors denote A_2 as B_{22}!) Both coefficients characterize the pairwise interactions between solute particles in dilute solutions, since each can be expressed in terms of the orientationally averaged potential of mean force $W(r)$ (assuming isotropic interactions). Namely [18],

$$B_2 = 2\pi \int_0^\infty [1 - \exp -(W/kT)] r^2 \, \mathrm{d}r. \tag{3.7}$$

As discussed in Chapter 4, $W(r)$ is essentially the interaction free energy between two protein molecules whose center-to-center distance is r, and the effects of the solvent are assumed to have been averaged out. It is important to realize that this averaging over the solvent molecules is seldom carried out explicitly. Rather, one usually assumes some approximate form for $W(r)$. In addition, one often treats the protein molecules as homogeneous, spherical particles, such that averaging over the possible orientations of two interacting proteins is not carried out. This is incorrect for realistic, anisotropic models of protein molecules; see e.g., refs. [19] and [20]. We discuss later an ambitious attempt by Lenhoff and colleagues to calculate the second virial coefficient for certain globular proteins, using realistic models for the protein structure, that takes into account the contribution of the orientational configurations. These are based on a generalization of Eq. (3.7) [19, 21, 22]:

$$B_2 = \frac{1}{16\pi^2} \int_0^{2\pi} \int_0^{\pi} \int_0^{2\pi} \int_0^{2\pi} \int_0^{\pi} \int_0^{\pi} [1 - \exp -(W/kT)] r^2 \times \mathrm{d}r \sin\theta \, \mathrm{d}\theta \, \mathrm{d}\phi \sin\beta \, \mathrm{d}\beta \;\, \mathrm{d}\gamma. \tag{3.8}$$

Here θ and ϕ denote the angular location of the second molecule relative to the first, and the orientation of the second molecule is specified by the Euler

angles α, β, and γ. The potential of mean force, W, is now a function of all these configurational variables, Ω, and r, where we denote the set of angular variables by Ω, following Lenhoff and colleagues [19]. One can divide the integral into two parts, one representing the excluded volume contribution and the other the contribution from the various positive and negative interactions between the proteins. Thus one has

$$B_2 = \frac{1}{16\pi^2}\int_\Omega \left\{\sigma^3/3 + \int_\sigma^\infty [1 - \exp -(W/kT)] r^2\,\mathrm{d}r\right\}\mathrm{d}\Omega. \tag{3.9}$$

Here W is infinite for $r < \sigma$, due to the hard core interactions (excluded volume effect).

3.4 Scattering theory

In addition to determining crystal structure, scattering techniques are widely used to probe the interactions between globular proteins. Globular proteins in solution scatter radiation, due to the local fluctuations in concentration that result from thermal energy. Experimental measurements of the intensity of scattered radiation, using light scattering, small angle X-ray scattering (SAXS) or neutron scattering provide a powerful method of probing the interactions between the protein molecules. As is well known, the intensity of scattered radiation, $I(q)$, is a function of the scattering wave vector q, where $q = (4\pi/\lambda)\sin(\theta/2)$, where θ is the scattering angle and λ is the wavelength of the incident radiation. The static (quasi-elastic) scattering intensity can be written as the product of two terms [23]:

$$I(c, q) = I(0, q)S(c, q), \tag{3.10}$$

where $I(0, q)$ is the form factor, $S(c, q)$ is the structure factor, and c is the protein concentration. The form factor describes the scattering from each individual protein molecule; it is the Fourier transform of the electron density contrast associated with the molecule. In an ideal solution in which there are no interactions between particles, the total scattering is just the sum of the scattering of the individual molecules. For non-ideal solutions, the structure factor accounts for the deviation from ideality and is the Fourier transform of the two-point particle distribution function, $g(\mathbf{r})$. In particular, $S(c, q)$ is the Fourier transform of the spherically averaged distribution function $g(r)$:

$$S(c, q) = 1 + 4\pi\rho\int [g(r) - 1]\left(\frac{\sin qr}{qr}\right) r^2\,\mathrm{d}r, \tag{3.11}$$

where, as above, $\rho = cN_\mathrm{A}/M$ is the number of particles per unit volume and c is the particle concentration ($\mathrm{g\,cm^{-3}}$).

The form factor is obtained from measuring the scattering intensity at low protein concentration. In SAXS it can be analyzed in terms of the radius of gyration of the molecule, R_g, if $2\pi R_g q < 1$, by using the Guinier approximation [24]:

$$I(c \to 0, q) = I(0,0) \exp\left(-\frac{R_g q^2}{3}\right). \tag{3.12}$$

From the plot of $\log I(c, q)$ as a function of q^2 for small $q \to 0$ (Guinier plot), one can determine $I(0, 0)$ and the slope, which yield the molecular weight and radius of gyration of the protein, respectively. Once this is known, measuring $I(c, q)$ at finite concentrations yields the structure factor.

The structure factor at the origin is related to the normalized osmotic compressibility [24]:

$$S(c,0) = \frac{RT}{M}\left(\frac{\partial \Pi}{\partial c}\right)^{-1}, \tag{3.13}$$

where R is the gas constant ($8.31\,\mathrm{J\,mol^{-1}\,K^{-1}}$), T is the absolute temperature in kelvin, and Π is the osmotic pressure of the protein solution. The equation of state for the osmotic pressure in terms of the concentration is given by Eq. (3.7). Therefore the value of the structure factor at the origin can be obtained from the following equation:

$$S(c,0) = \frac{I(c,0)}{I(0,0)} = \frac{1}{1 + 2MA_2c + \cdots} = \frac{1}{1 + 2B_2\rho + \cdots}. \tag{3.14}$$

In the case in which $2MA_2c \ll 1$, one can thus obtain an expansion of $S(c, 0)$ in powers of the concentration, from which one can determine the second virial coefficient:

$$\frac{1}{S(c,0)} = 1 + 2MA_2c + \cdots. \tag{3.15}$$

This is a very useful expression. Note that if A_2 is positive, the net interaction between molecules is repulsive so the structure factor at the origin, $S(q = 0)$, is less than unity. If A_2 is negative, the net interaction is attractive and $S(q = 0)$ is greater than unity.

In light scattering, one usually analyzes the data in terms of the Rayleigh ratio, R_θ. This is the excess scattering intensity per unit volume and solid angle, normalized by the incident intensity, and is given by

$$R_\theta = KMcP(q)S(q), \tag{3.16}$$