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 supersaturated 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

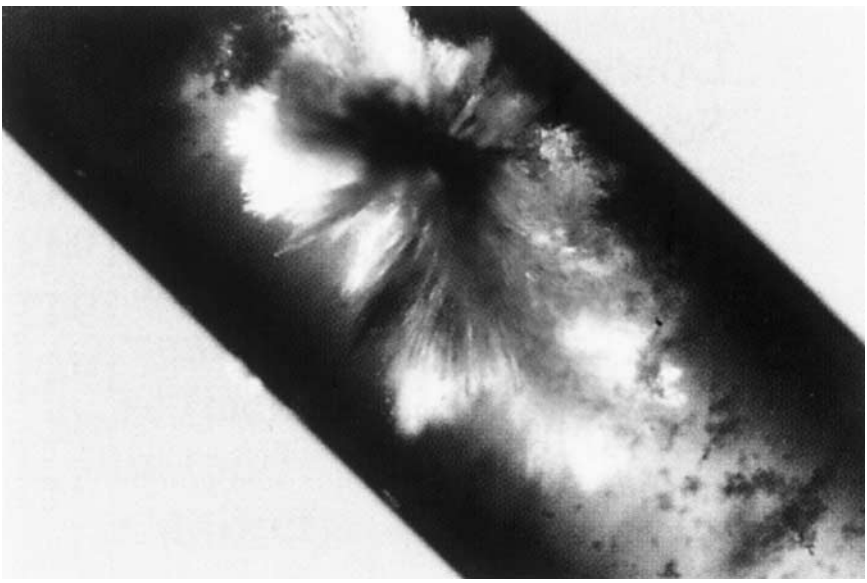

Figure 1.1. Fine needle crystals of hemoglobin, much like those grown by Hünfeld in 1840. Reprinted, with permission, from ref. [10]. For a color image, please see Plate Section.

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.

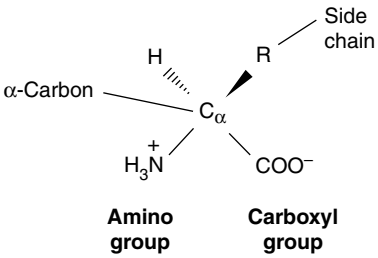

Figure 2.1. Anatomy of an amino acid. Except for proline and its derivatives, all of the amino acids commonly found in proteins possess this type of structure. From ref. [11]. Reprinted with permission of Brooks/Cole, a division of Thomson Learning: www.thomsonrights.com.

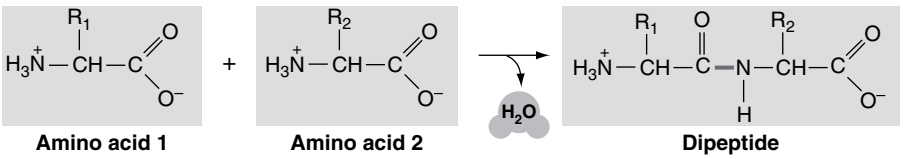

Figure 2.2. Peptide formation is the creation of an amide bond between the carboxyl group of one amino acid and the amino group of another amino acid; R_1 and R_2 represent the R groups of two different amino acids. From ref. [11]. Reprinted with permission of Brooks/Cole, a division of Thomson Learning: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