1

Introduction

1.1 Life

We are surrounded by microbes, plants and animals that we immediately recognize as living beings (Figure 1.1). However, it is still difficult to provide a concise definition what life is. Perhaps the most useful definition for this book is that life is a unit capable of chemical activities, which can reproduce and evolve.

Chemical activities, which involve conversions of energy and matter, are called *metabolism*. These activities capture energy and chemical matter in different forms. Thousands of chemical activities take place simultaneously in a living organism and they must be well coordinated or regulated to maintain the stability of the living unit.

Reproduction of the unit (generating new units) provides both the continuity and the variation that is also an important characteristic of life. The combination of reproduction, horizontal transfer of information and "erroneous" duplicates provides the basis of evolution. In other words, the composition of the unit should be able to change over time to better adapt to the changing environmental conditions. Living organisms appear in very different forms and follow very different life-styles. However, the basic characteristics of life (including metabolism, reproduction and evolution) are provided and governed by very similar sub-structures: biological macromolecules and cells.

1.2 Levels of Organization of Life

The living world has several hierarchical levels, ordered from the smallest to the largest. At the bottom are molecules, a mix of inorganic and organic compounds and biological

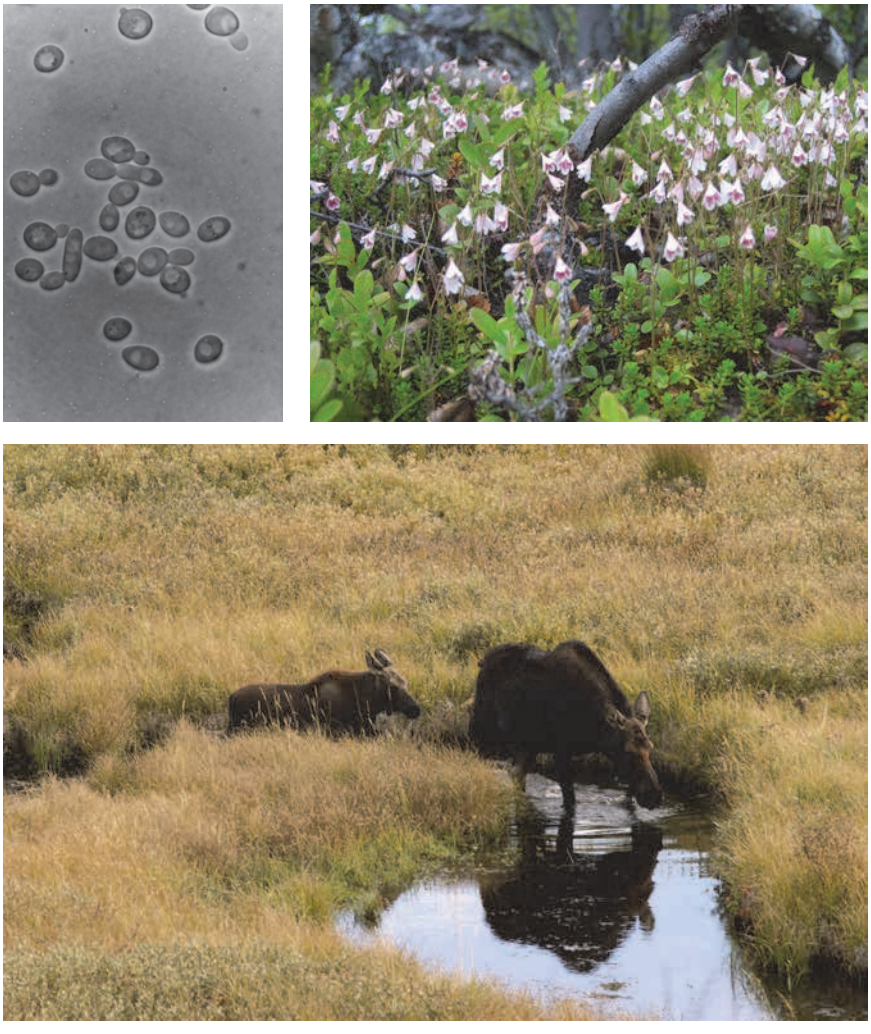

Fig. 1.1 ■ Living organisms are found in numerous different forms. *Left*: A microscope picture of baker's yeast (*Saccharomyces cerevisiae*) cells (by the courtesy of Concetta Compagno). *Right*: Linneas (*Linnea borealis*) covering vast areas of Lapland (by the courtesy of Bernarda Rotar) and *bottom:* moose, the largest land animals in Scandinavia (by the courtesy of Aca.Pixus.dk). Within these different macro-forms very similar molecular structures can be found, which determine the form and lifestyle of the carrier organisms.

macromolecules, followed by sub-cellular structures, cells, tissues, organs, organisms, populations, communities and the biosphere, which encompasses all biological communities on the Earth.

Macromolecules are central in all living organisms. They are giant polymers consisting of repeating units. These repeating units may or may not be identical, and are connected with covalent or non-covalent bonds. Macromolecules perform a multitude of functions, which are the basis of metabolism, reproduction and evolution, such as energy or information storage, reaction catalysis, coordination and regulation, communication, structural

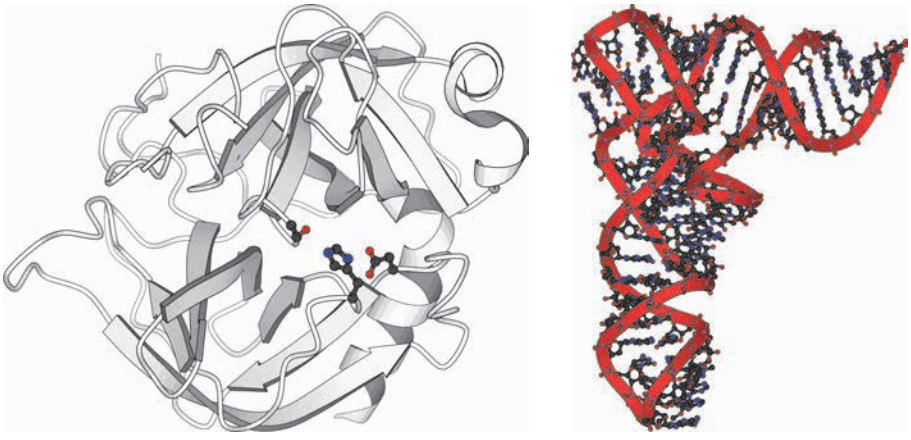

Fig. 1.2 ■ A simplified picture of two bio-macromolecules, which are the focus of our further chapters. *Left:* The structure of a well-known protein, chymotrypsin (PDB: 4CHA). *Right:* A nucleic acid molecule, yeast tRNAPhe (PDB: 1EHZ).

support, defense, movement and transport. On the basis of chemical composition we talk about three different kinds of macromolecules: (i) peptides and proteins that are polymers of amino acid residues, (ii) nucleic acids, which are polymers of nucleotides, and (iii) carbohydrates, which are polymers of sugars (Figure 1.2). Other central molecules that should be mentioned here are the lipids. Although they are not macromolecules, they self-assemble into large aggregates of macromolecular dimensions, including the lipid bilayer (an important building block of cell membranes), micellar aggregates containing bile molecules, and the aggregates of lipoproteins that transport cholesterol and fat in the blood stream. In the following chapters we will try to understand the structures of bio-macromolecules and link them to their functions and the higher levels of the living world.

The basic unit of life is a *cell* (Figure 1.3). Cells are surrounded by a plasma membrane, which separates each cell from the external environment and creates a segregated compartment with a controlled internal environment. Cells show two organizational patterns: (i) prokaryotic, characteristic for Bacteria and Archaea, and (ii) eukaryotic, characteristic for Eukarya. Prokaryotic cells usually exist as single cells and are smaller than eukaryotic ones, typically on the order of 1 μm in diameter. The basic structure of a prokaryotic cell is defined by a cellular membrane, an intracellular nucleoid containing DNA, and the cytosol holding the rest of the intracellular material, where ribosomes, enzymes and cytoskeletal elements are found. Eukaryotic cells are usually at least ten times larger than prokaryotic cells and more complex, with inner membranes separating compartments and organelles. The organelles include: (i) the nucleus, storing genetic material and the replication and gene transcription systems (ii) the cytosol, where protein synthesis and many essential biochemical reactions take place, (iii) the mitochondrion, a power plant and energy storage compartment, (iv) the endoplasmatic reticulum and Golgi apparatus, where proteins are matured and sorted to further locations, (v) the lysosomes or vacuoles, where polymeric macromolecules, such as proteins, are recycled into usable metabolites.

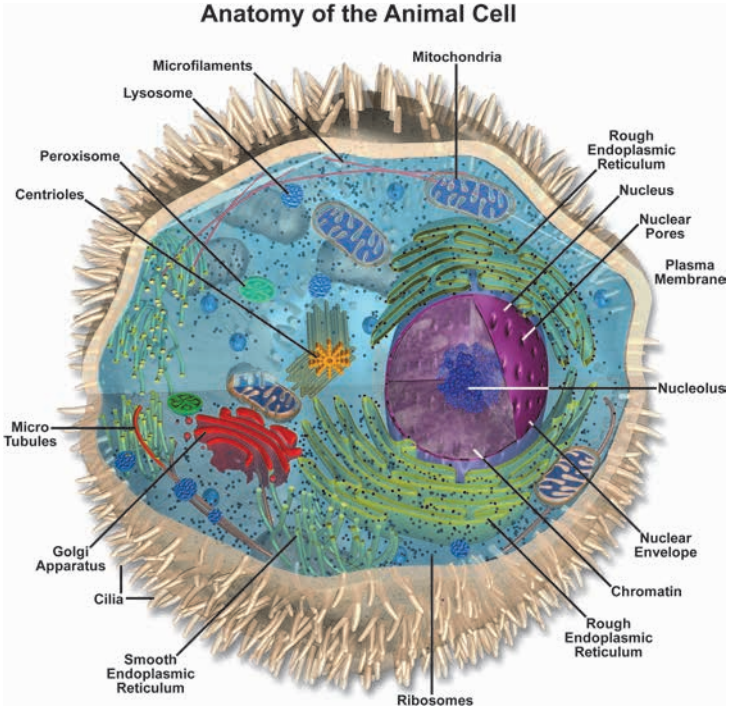

Fig. 1.3 ■ A schematic picture of an animal cell showing sub-cellular structures, such as nucleus, membrane systems (ER), mitochondrion, etc. (Made by Michael W. Davidson, Florida State University.)

All organisms on Earth seem to originate from a single unicellular *organism*. The main reasons why one can claim that all organisms originate from the same cell is that not only do all living species use the same nucleotides and amino acids despite many other possibilities, but the genetic code (the dictionary for translation from the language of nucleic acids to the one of proteins) is the same. In addition, central molecular systems like transcription and translation are strongly related. A smaller molecule like ATP is the universal currency of energy in all living organisms, although in principle many other choices would have been possible.

Today, many millions of different organisms that do not interbreed with each other are found and we call them species (Figure 1.4). They are all adapted to their different environments and in a naive sense they may seem perfect. However, a particular life form may not be fit tomorrow and thereby become extinct, like so many other species in the past, which have previously populated Earth. Due to changes of environment, new and better-fit species constantly evolve over time, and this evolution works by gradually changing the structures of macromolecules.

The unfolding of events leading to the present diversity is expressed as an evolutionary tree showing the order in which species split and evolved into new species. This tree traces the descendants coming from ancestors that lived at different times in the past.

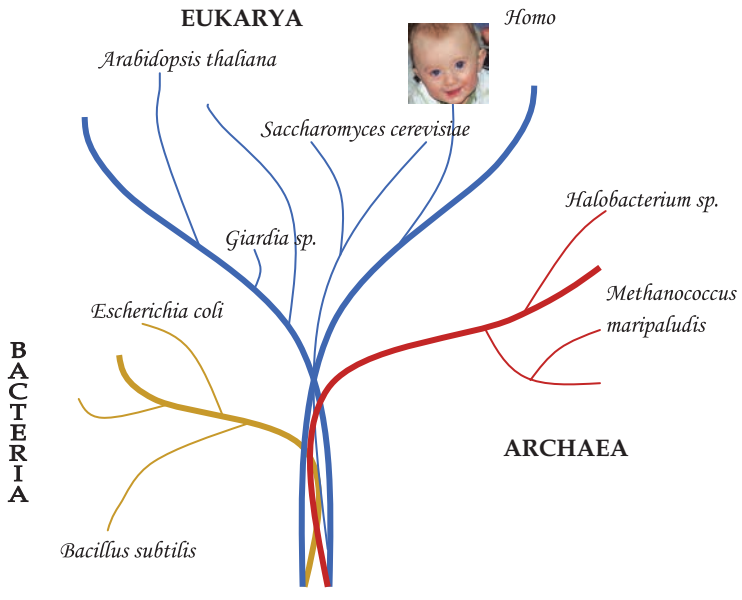

Fig. 1.4 ■ A simplified tree of life. The common progenitor originated approximately 4 billion years ago. The position of the first branchings, occurring between the progenitor of Bacteria, Archaea and Eukarya, are still unclear.

In other words, the evolutionary tree shows the evolutionary relationship among modern and ancient species. It is important to understand the evolutionary relationship between organisms when one compares the structure of macromolecules from these organisms, being involved in similar processes. Some of the earlier branching is difficult to reconstruct because there are no available fossils. However, based on molecular evidence in modern organisms, we can separate all living organisms into three domains, which have been evolving separately for more than 1 billion years: (i) Archaea, (ii) Bacteria, and (iii) Eukarya. Even if they superficially look similar, Archaea and Bacteria separated into distinct lineages very early during evolutionary history.

Archaea are often inhabitants of extreme environments, such as hot and acidic springs, sea depths and salt brines, but can also be found in more "normal" environments. Their replication, transcription and translation machinery resembles the eukaryotic machinery, while their metabolism and energy conversion resemble the bacterial ones.

Bacteria consist of more than a dozen sub-groups, also called clades, but the most important are: Protobacteria, Cyanobacteria, Spirochetes, Chlamydias and Firmicutes. The Protobacteria are the largest and a very diverse group, including one of the best-studied organisms, *Escherichia coli*. Sometimes bacteria are divided, on the basis of their cell wall composition, into gram-positive, including *Bacillus subtilis*, and gram-negative, including *E. coli*. Bacteria exhibit the greatest biochemical diversity.

Eukarya can be divided into four groups: Protista, Plantae, Fungi and Animalia. The Protista contain mostly single celled organisms and have a polyphyletic origin, meaning

that some represent very primitive eukaryotes, such as *Giardia*, while some are closely related to animals, such as *Dictyostelium*, or plants, such as red algae. Phagotrophy, a feeding mode to form a pocket in the plasma membrane and enclose the "food", is a hallmark of Eukarya.

1.3 Short History of Life on Earth

The theory of chemical evolution holds that conditions on the primitive Earth, around 4 billion years ago, led to the emergence of the first biological molecules. Oparin and Haldane independently suggested in the 1920s that if Earth's first atmosphere was reducing, and if there was a supply of external energy then a range of organic compounds might be synthesized. In the 1950s, Stanley Miller and Harold Urey mimicked these conditions in the lab. Water vapor, hydrogen gas, ammonia and methane gas were exposed to sparks, and after a few days the system contained several complex molecules, such as amino acids and nucleic acid bases, the building blocks of today's life. When the monomeric units were present, it was not so difficult to achieve polymerization even under abiotic conditions. However, how could the first peptides and nucleic acids become "alive"? In other words, how could they start reproducing and evolving?

The term *replicator* means a structure that can arise only if there is a preexisting structure of the same kind in the vicinity. For example, a supersaturated solution crystallizes if a small seeding crystal is added. However, this represents a simple replicator relying on a single structure. More sophisticated replicators could exist in several forms and thereby could have contributed to heredity. For sustained evolution, an indefinite number of forms and indefinite variation in heredity is necessary. The first artificial replicator, a simple hexadeoxynucleotide not needing enzymes for its replication (polymerization from the present mono-units) was synthesized by von Kiedrowski in 1986. The first short RNA molecules may have had the ability to catalyze the polymerization of offspring molecules. The first replicating RNA molecules competed successfully with their own erroneous copies and with other less-efficient systems for the monomers needed for their replication. Even if self-replicating RNA molecules fulfill the above criteria for life, the path to the first cells was still more sophisticated. One of the main following steps was to include peptides and proteins to establish the RNA — protein world, followed by the introduction of membrane systems, thereby segregating the primitive cell from its environment.

The origin of the first cell, the common progenitor of all living organisms, could be approximately 3.5 billion years ago, and at that time simple replication and translation machineries already existed. One hypothesis suggests that, during the following 2 billion years, the unicellular system evolved to represent a fine net of metabolic reactions connected to increasingly sophisticated machineries for nucleic acid replication and RNA to protein translation, also keeping plasticity, enabling the cell to respond to the demands of

the ever-changing environment. During this period, the first living cells were still dependent on organic compounds, which were the primary source of energy, and had abiotic origin. Later, approximately 2.5 billion years ago, one of the major steps was the evolution of the ability to use the energy of sunlight to power metabolism. Photosynthesis provided energetic independence and soon resulted in vast quantities of organic materials and oxygen. The evolution of aerobic metabolism significantly changed cellular biochemistry. Many enzymatic reactions became dependent, directly or indirectly, on the presence of oxygen. Aerobic metabolism allowed cells to grow larger. Some of the further major transitions include the origin of sex, the origin of multicellular organisms and the origin of social groups. Behind all these events stood proteins, nucleic acids, carbohydrates and lipids, with their evolving structures and functions.

1.4 What is Structural Biology and When Did It Start?

The field of structural biology focuses on a classical insight: in order to understand, we need to see. "Seeing is believing", or "a picture says more than a thousand words" are well-known phrases. This is true whether we deal with large objects, as in astronomy and astrophysics, medium-sized objects such as birds or fishes, or with very small objects like biochemical systems or particle physics. Structural biology is the science that tries to make the sub-cellular and molecular objects of biology visible and understandable.

It is difficult to identify the very beginning of structural biology. One important step is the purification of the fundamental components. Friedrich Miescher discovered and isolated DNA in 1869. The understanding of the biological role of DNA did not start until 1944 when Avery, MacLeod and McCarty showed that DNA is the genetic material. Elucidation of the structure of DNA in 1953 was a major milestone in structural biology. Francis Crick and James Watson, using diffraction data obtained by Rosalind Franklin and Maurice Wilkins, deduced a model for DNA. This model led to a detailed insight of the replication of DNA, the transcription of DNA to RNA and also the key steps of translation, central activities in molecular biology.

In a review from 1964, James Watson expressed: "Unfortunately, we cannot accurately describe at the chemical level how a molecule functions unless we know first its structure." This describes the situation in a nutshell and it is exemplified in one field of biology after the other.

Proteins have long been known, but the molecular nature of them was poorly understood. Jöns Jacob Berzelius (1779–1848), the well-known Swedish chemist, introduced the term protein. Proteins were classified as colloids without defined structures and shapes. The first crystallization of proteins may have been of hemoglobin, in 1840 by Hünefeld.

At this time, the crystals were called "blood crystals" and it was not realized that the red crystals were built of a protein. Several other proteins were also crystallized early on. The nature of proteins became better understood when Theodor (The) Svedberg could show with his ultracentrifuge that proteins have unique molecular weights.

During the 19th century the action of gastric juice on the degradation of solid proteins was thoroughly studied. One active ingredient was called pepsin but its nature was entirely unclear. Gradually, the catalytic substance was given names as "ferment" or "enzyme". Enzymes were believed by (among others) the Nobel laureate Willstätter to be of a different nature than lipids, carbohydrates or proteins and present in only very low concentrations in plants or animals. J.B. Sumner and subsequently J.H. Northrop showed that enzymes are proteins with unique structures, since urease and pepsin could be purified and crystallized. John D. Bernal and Dorothy Crowfoot (later Hodgkin) could show that pepsin crystals diffracted X-rays when kept in a moist environment, thus demonstrating that proteins would have a specific structure, which was lost if dried out. In the same period, F.C. Bawden, N.W. Pirie and W.M. Stanley crystallized a number of viruses. The perfection of crystallization and the crystallographic analysis of protein and enzyme crystals took several decades until it matured in the well resolved crystallographic structures of myoglobin and hemoglobin, in 1959 and 1968, respectively.

Structural biology includes a number of methods in addition to diffraction and scattering methods. In an early phase, electron microscopy was already an important technique to obtain an insight into the organization of biological systems and macromolecules. One major advance was the analysis of virus particles by Caspar and Klug in the end of the 1950s and beginning of the 1960s. The symmetry principles could be deduced, and for the larger viruses different functional components were identified. Another development in the field of electron microscopy was the studies of 2D crystals of bacteriorhodopsin studied by R. Henderson and N. Unwin using electron diffraction. This opened new possibilities, but only a limited range of objects yielded material good enough for structural studies. Subsequent to these developments, the single particle reconstruction studies of large molecular complexes at cryo-temperatures (cryo-EM) and tomography have very significantly extended the capabilities of electron microscopy to contribute entirely new insights of structural biology at a range of resolutions. The single particle reconstruction has become a new revolution in the field, with a capacity to analyze complexes that were poorly or not at all available before at high resolution. In addition, cryo-EM has the capacity to identify several different conformations in a single sample, adding insights into the dynamics of functional molecules.

Structural biology has moved numerous systems from an understanding where the molecules are represented by blobs to where the atomic coordinates are available, as well as details of molecular interactions.

A limitation of crystallography is that it gives still pictures of the molecular systems studied, with only limited insight into dynamics. In fortunate cases, a number of states can be crystallized and characterized at atomic resolution. However, in many cases one

would need insights into states that are not accessible to crystallization, maybe because they are too short-lived, to understand the dynamics of the system. Here, NMR-studies can sometimes give the information that is missing.

Generally, NMR spectroscopy can provide structures, which are particularly valuable when crystals cannot be obtained. When both NMR and crystallographic structures are available, the quality of the crystallographic information is normally better. However, the unique contributions of NMR spectroscopy come from dynamic studies of systems where the structures are already known. Here, the mobility and details of transient interactions can be characterized.

To obtain optimal information several methods should be employed. Wrong or partial information can be corrected or extended. In the best-understood systems, physical and theoretical chemists have contributed their experimental or computational methods to get additional angles on the understanding of the system.

1.5 A Short Summary of the Book

This book is a textbook of structural biology for undergraduate and graduate students. The focus is to cover the central and most interesting aspects of structure, combined with a focus on interesting biology. The book makes no attempt to cover the entire fields of biology or molecular biology. One selection principle is a focus on systems where we know the structures reasonably well.

We try to provide a comprehensive coverage of the structural and functional understanding of proteins, nucleic acids, lipids and membranes as well as carbohydrates. The book makes no attempt at describing the methods used to obtain the results. This would require a separate volume. In addition to the basic structural knowledge of protein, nucleic acid, lipids and carbohydrates there is a significant coverage of the steps involved in the expression of the genetic information in DNA into proteins. Likewise, the breakdown of macromolecules is covered. Much of biology is related to membranes. They enclose cells or cellular compartments. The passage of material and information across membranes is crucial for all cellular biology and structural insights are rapidly increasing. In multicellular organisms, cell-cell contacts and interactions are essential for coordinated activities. In relation to this, the insights into how cells and organisms move is increasingly better understood. These fields are described in the text.

Insights into the evolution of biological systems and functional genomics also benefit from structural studies. The rate by which the DNA sequences of complete genomes are produced has generated an enormous database that is rapidly growing. DNA or protein sequences can confidently be identified as long as the sequence identity is reasonably good. In many cases, the sequences themselves are insufficient. However, the structural

relationship can both suggest the evolutionary relationship of a protein as well as its function. The final chapter gives some approaches to use various predictive methods to access structural insights from sequence data alone.

Further Reading

Avery GT, Macleod CM, McCarty M. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonuleic acid fraction isolated from pneumococcus type iii. *J Exp Med* **79**: 137–158.

Bernal JD, Crowffot DC. (1934) X-ray photographs of crystalline pepsin. *Nature* **133**: 794–795.

Franklin RE, Gosling RG. (1953) Molecular configuration in sodium thymonucleate. *Nature* **171**: 740–741.

Fruton JS. (2002) A history of pepsin and related enzymes. *Quart Rev Biol* **77**: 127–147.

Kendrew JC, Dickerson RE, Strandberg BE, *et al*. (1960) Structure of myoglobin: a thre-dimensional Fourier synthesis at 2 A resolution. *Nature* **185**: 442–427.

McPherson A. (1991) A brief story of crystal growth. *J Cryst Growth* **110**: 1–10.

Miescher F. (1871) Über die chemische Zusammensetzung der Eiterzellen. *Hoppe-Seyler's medicinisch-chemische Untersuchungen* **4**: 441–460.

Northrop JH. (1929) Crystalline pepsin, *Science* **69**: 580.

Olofsson I, Liljas A, Lidin S. (2014) From a grain of salt to the ribosome. The history of crystallography as seen through the lens of the Nobel Prize. World Scientific, Singapore.

Sumner JB. (1926) The isolation and crystallization of the enzyme urease: preliminary paper. *J Biol Chem* **69**: 435–441.

Watson JD, Crick FH. (1953) Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* **171**: 737–738.