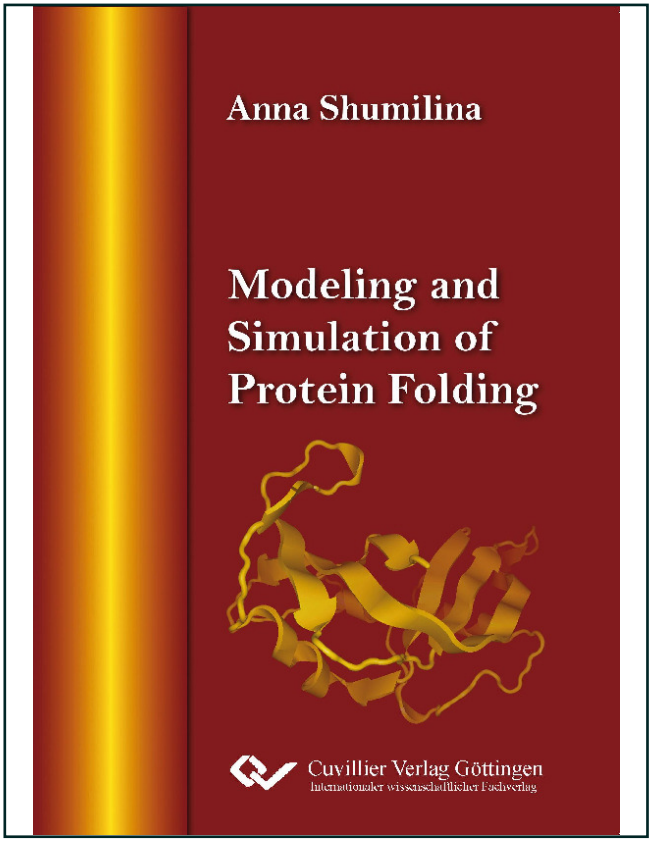
Anna Shumilina (Autor)

# Modeling and Simulation of Protein Folding




https://cuvillier.de/de/shop/publications/325

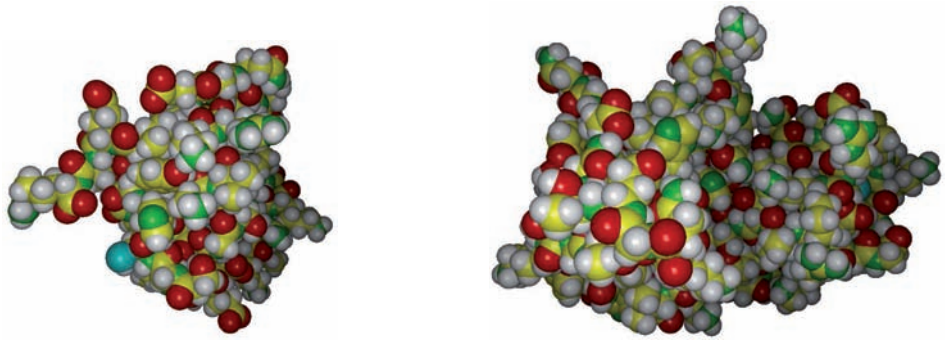
# Chapter 1

# Biological, Chemical, and Physical Background

## 1.1 Introduction

Proteins are essential components of any living cell. They have very diverse functions: catalyze chemical reactions and control gene expression, constitute a cytoskeleton and perform muscle contraction, transport electrons, ions and uncharged molecules, enable recognition of cellular signals or alien invasion. The properties of a protein molecule are determined by its spatial structure (see Fig. 1.1) and location of charged atom groups, which often have to be very specific for a protein to perform a certain function.

The spatial structure of a protein depends on its chemical composition. A protein consists of one or more associated *polypeptide* chains, which are built of consequently



(a) (b)

Fig. 1.1: Space-filling representations* of two native protein structures determined by nuclear magnetic resonance spectroscopy†. (a) Glucocorticoid receptor DNA-binding domain. (b) Bovine pancreatic ribonuclease A.

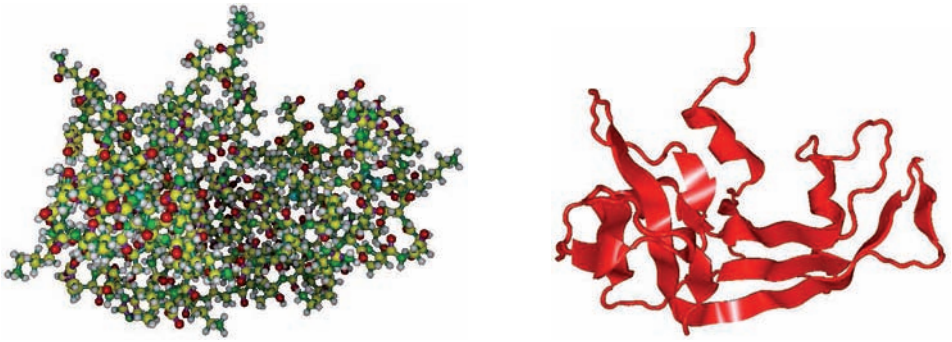
*Molecular images here and further in the text are generated by the program SiViProF, developed in course of this work. For an explanation of the color notations see Table A.1 in Appendix A.

†Atomic coordinates are obtained from RCSB Protein Data Bank (see Section 1.10 for details), from records 1GDC by Baumann *et al.* [6] and 2AAS by Santoro *et al.* [7].

connected amino acid residues (see Section 1.2 for details). The number of residues can vary from about fifty to many thousands, depending on protein functions. Shorter sequences are usually referred as *peptides*[*] and often do not have any fixed spatial arrangement in solution.

The sequence of residues is unique for each protein and believed to predetermine the result of folding of the synthesized chain into its native state in a proper environment. The latter statement is named *Anfinsen's dogma* after the Nobel prize laureate Christian B. Anfinsen, who has shown in 1961 that ribonuclease (see Figures 1.1 (b) and 1.2 (a, b)) with reduced disulfide bonds and disrupted tertiary structure was able to restore its enzymatic activity upon removal of the denaturating agent [8]. Later it was proven that also many other small proteins are able to refold *in vitro*[†] into their functional form. These results motivated numerous attempts to compute native three-dimensional structure of proteins based on given amino acid sequences.

Protein structure prediction is a subject of intense research, given the importance of its academic and medical applications. A large number of known protein sequences is already available, and the amount of this data grows rapidly. By contrast, an experimental determination of protein three-dimensional structures by means of X-ray crystallography or nuclear magnetic resonance spectroscopy is relatively expensive and time consuming (see Subsections 1.9.1 and 1.9.2). Computation of native protein structures from their amino acid sequences could contribute to the understanding of the organization of living organisms on the molecular level and give a clue to treatment of many diseases. It would also enable more rational drug design, helping to lower the costs and the amount of time required to introduce new medications.



(a) (b)

FIG. 1.2: Some more insight into the structure of bovine pancreatic ribonuclease A[‡]: (a) the ball-and-stick model visualizing all atoms and bonds between them, (b) the ribbon model showing the fold of the main chain. A discussion about different models for protein visualization follows further in the text, see Subsections 1.5.4 and 1.6.2.

[*]Peptides consisting of two, three, or a few amino acid residues are called *dipeptides*, *tripeptides* or *oligopeptides* respectively.

[†]Outside a living organism, literally, *in glass* (Latin).

[‡]Atomic coordinates are the same as in Figure 1.1 (b).

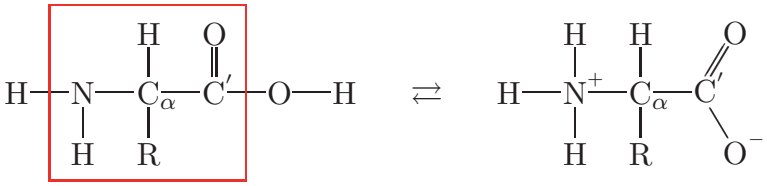


Fig. 1.3: General structure of amino acids in a nonionized and an ionized form. R stands for a side chain. The red box outlines a non-terminal residue after inclusion into a polypeptide.

## 1.2 Chemical structure of proteins

There are 20 different amino acids that are used by cells as building blocks in protein synthesis. They are listed in Tables 1.1-1.3. All of them, with exception of proline, have a common part, containing an *amino* ($-NH_2$) and a *carboxyl* ($-COOH$) group (Fig. 1.3). The carbon atom of the carboxyl group is conventionally marked as $C'$ in order to distinguish it from the $\alpha$-carbon bonded to the amino group. The distinctive part of an amino acid is called the *side chain*. Its non-hydrogen atoms are labeled using subsequent Greek letters, starting from the atom linked to $C_\alpha$. In case of branching, letters are additionally supplied by indexes (see Table A.2 in Appendix A for details).

In course of protein synthesis the common fragments of amino acids are joined together by *peptide bonds*, thereby constituting the *main chain*, or the *protein backbone* (Fig. 1.4). As a result of peptide bond formation, the hydroxyl group ($-OH$) at $C'$ and a hydrogen from the amino group of the next amino acid are removed. Although the structure of proline is somewhat different, it allows its molecules to be incorporated into a chain in a similar way (see Subsection 1.6.1).

The residues are appended to the carboxyl end in a certain order, prescribed by the corresponding genetic code. This procedure is termed *translation*. The details of the protein synthesis that are relevant for initial arrangement of atoms in a nascent protein are discussed in Section 1.8. After completion of translation, some chemical alternations of standard amino acids can be performed as a part of a controlled process, termed *posttranslational modification*.

The sequence of residues in a protein is called its *primary structure*. It is written using conventional one- or three-letter abbreviations (see Table A.2 in Appendix A), starting from the amino end. The reverse order of residues corresponds to another protein.

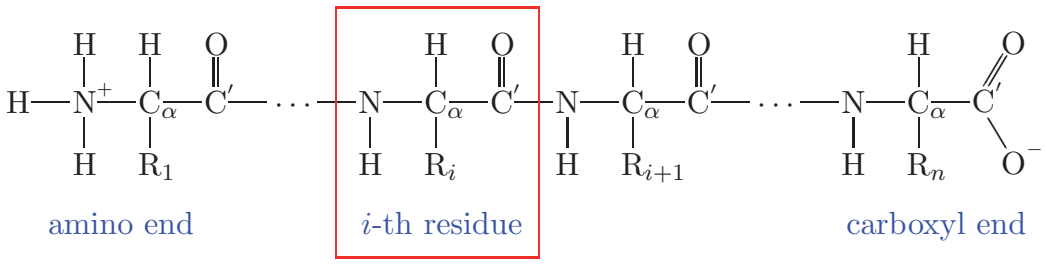
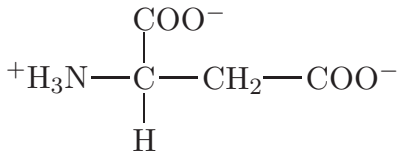
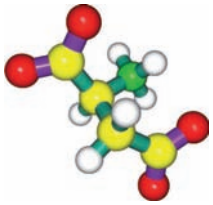
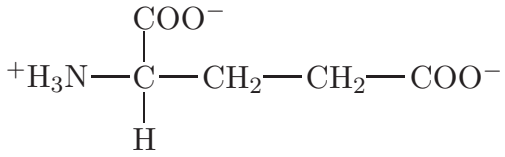
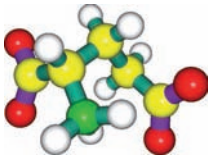
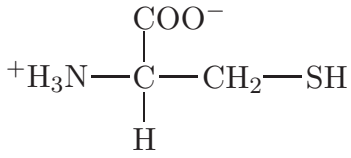
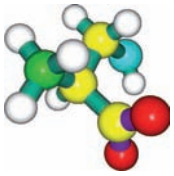
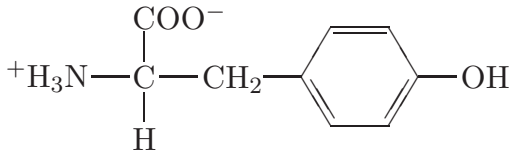
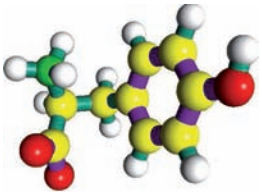
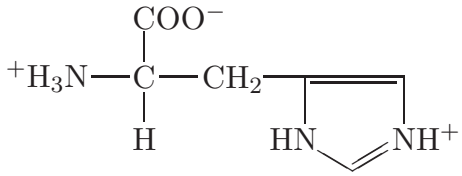
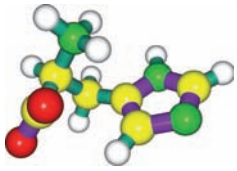
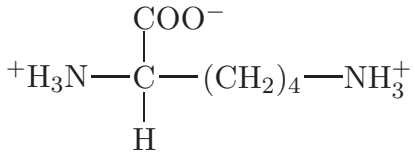
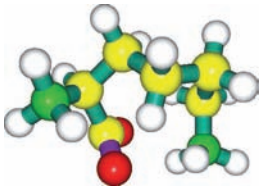
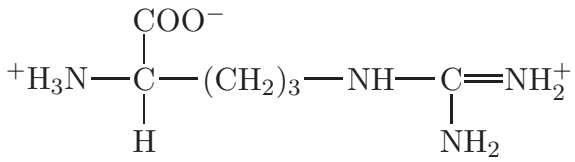
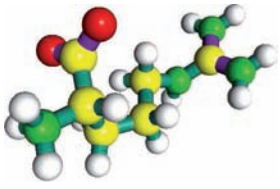



Fig. 1.4: A polypeptide chain. Peptide bonds connect $C'$ and N atoms.

Once or even before the polypeptide chain is completely synthesized, it adopts a certain conformation, which is responsible for the protein functions. The folding pathway and the resulting structure are largely determined by the properties of the constituent amino acid residues. For example, if folding happens in cytosol, which represents mainly a mixture of water with salts, nonpolar side chains seek to avoid

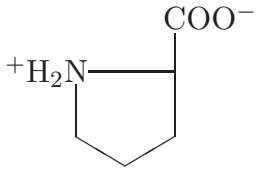
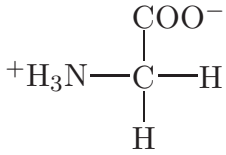
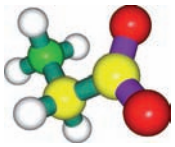
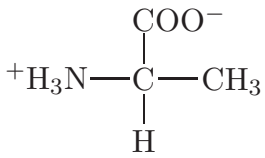
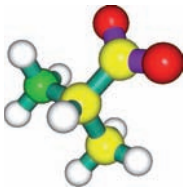
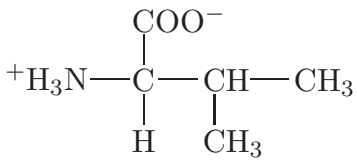
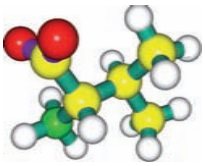
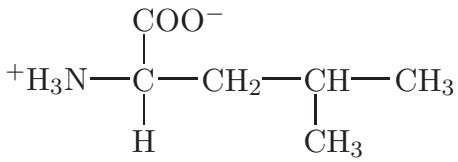
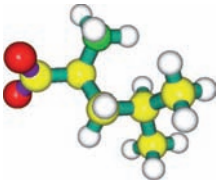
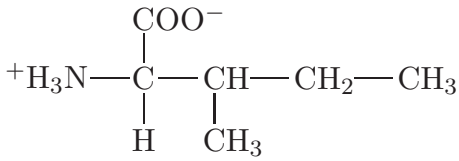
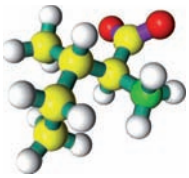
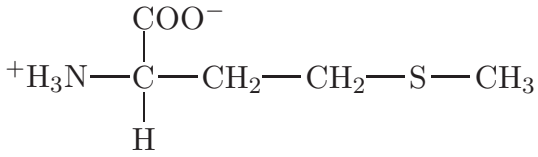
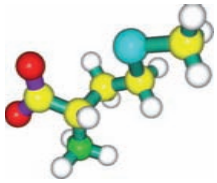
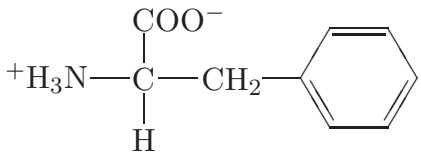
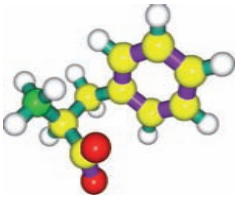
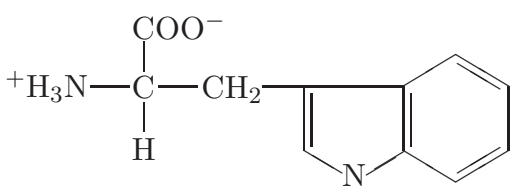
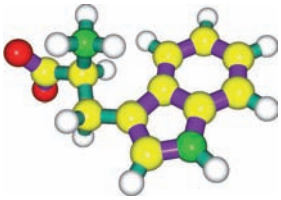
TABLE 1.1: Amino acids with hydrophilic ionizable side chains.

| **Name** | **Chemical formula*** | **Molecular structure†** |
|---|---|---|
| Aspartic acid |  |  |
| Glutamic acid |  |  |
| Cysteine |  |  |
| Tyrosine |  |  |
| Histidine |  |  |
| Lysine |  |  |
| Arginine |  |  |

*The presented form considerably prevails in physiological conditions, except for histidine: only about one fourth of histidine side chains is protonated in cytosol (see Section 1.4).

†In the last column the histidine side chain is depicted in the more probable non-protonated form. Single and double bonds in conjugated systems are treated as bonds having partial double character.

Table 1.2: Amino acids with hydrophobic side chains.

| Name | Chemical formula | Molecular structure |
|---|---|---|
| Proline |  |  |
| Glycine |  |  |
| Alanine |  |  |
| Valine |  |  |
| Leucine |  |  |
| Isoleucine |  |  |
| Methionine |  |  |
| Phenylalanine |  |  |
| Tryptophan |  |  |