1 Introduction

1.1 Relevance of biomolecular research

More than 30 million people, infected by the *human immunodeficiency virus* (HIV), suffer from the *acquired immune deficiency syndrome* (AIDS);[1] 2 billion humans carry the *hepatitis B* virus (HBV) within themselves, and in more than 350 million cases the liver disease caused by the HBV is chronic and, therefore, currently incurable.[2] These are only two examples of worldwide epidemics due to virus infections. Viruses typically consist of a compactly folded nucleic acid (single- or double-stranded RNA or DNA) encapsulated by a protein hull. Proteins in the hull are responsible for the fusion of the virus with a host cell. Virus replication, by DNA and RNA polymerase in the cell nucleus and protein synthesis in the ribosome, is only possible in a host cell. Since regular cell processes are disturbed by the virus infection, serious damage or even the destruction of the fine-tuned functional network within a biological organism can be the consequence.

Another class of diseases is due to structural changes of proteins mediated by other molecules, so-called prions. As there is a strong causal connection between the three-dimensional structure of a protein and its biological function, refolding can cause the loss of functionality. A possible consequence is the death of cells. Examples for prion diseases in the brain are *bovine spongiform encephalopathy* (BSE)[3] and its human form *Creutzfeldt–Jakob disease* (CJD).[4]

A further source for damaging cellular networks is protein misfolding followed by amyloid aggregation. In the case of *Alzheimer's disease* (AD), which is a neurodegenerative disease and the most common type of *dementia*, amyloid beta (Aβ) peptides in sufficiently high concentration experience structural changes and tend to form aggregates. Following the amyloid hypothesis, it is believed that these aggregates (which can also take fibrillar forms) are neurotoxic, i.e., they are able to fuse into cell membranes of neurons and open calcium ion channels. It is known that extracellular Ca^{2+} ions intruding into a neuron can promote its degeneration. About 24 million, mainly elderly, people are currently affected.[5]

[1] UNAIDS/WHO *AIDS Epidemic Update: December 2007*.
[2] WHO Fact Sheet No. 204 (2000).
[3] The economic impact of BSE is disastrous. Following *USA Today* from August 4, 2006, US beef exports declined from $3.8 billion in 2003, before the first mad cow was detected in the USA, to $1.4 billion in 2005.
[4] The WHO Fact Sheet No. 180 (2002) reports a rate of one per million people.
[5] Alzheimer's Disease International, Global Perspective **16**(3), 1 (2006).

This exemplified collection of diseases manifests the extraordinary importance of molecular, biologically motivated, research. It comprises the understanding of the synthesis of biomolecules such as proteins based on the genetic code (gene expression), the characteristics and specificity of folding and aggregation processes, as well as the unraveling of the dynamic and kinetic aspects of biological processes. In almost all such processes functional proteins are involved. Of course, treating patients suffering from these diseases is a *medical* problem, but revealing the nature of mechanisms behind the functioning of living organisms is an interdisciplinary task for the whole ensemble of natural sciences.

This is also the case when it comes to nanotechnological applications. Science and technology meet and partly fuse at smallest length scales. In particular, learning from biological systems – from complex networks of biological functionality or from structural properties of single molecules – has become more and more important and relevant for special-purpose applications on micrometer or even nanometer scales. This also includes the interaction of soft materials with solid matter, where systematic research is just in its early stages. Examples for such hybrid systems with enormous potential for future applications are, among many others, biosensors in the form of adhesion-specific nanoarrays for the identification of proteins in solution and also nanoelectronic circuits on polymer basis. On the experimental side, the progress in the development of high-resolution equipment and new experimental techniques not only allows for detection of what is happening on atomic scales, but also enables local manipulation of molecules which is essential for the design of specific applications. On the other hand, computational capacities have reached a level that now makes it possible to study macromolecular systems more systematically in simulations of suitable models by means of sophisticated numerical methods. Simulations are particularly relevant for investigations of processes that are currently still inaccessible to experimental research.

At this point, the challenge for theoretical physics, in particular, is twofold: first, the modeling and analysis of specific molecular structures at atomic scales and second, the generalization of the conformational (pseudophase) transitions accompanying structure formations processes within a mesoscopic frame. Both approaches facilitate the systematic understanding of molecular processes that is typically difficult to achieve in experiments.

Protein folding, peptide aggregation, polymer collapse, crystallization, and adsorption of polymers and proteins to nanoparticles and solid substrates have an essential feature in common: in all these processes, structure formation is guided by a collective, cooperative behavior of the molecular subunits lining up to build chain-like macromolecules. In this process, polymers and proteins experience conformational transitions related to thermodynamic phase transitions. For chains of finite length, an important difference of crossovers between conformational (pseudo)phases is, however, that these transitions are typically rather smooth processes, i.e., thermodynamic activity is not necessarily signalized by strong entropic or energetic fluctuations. The interest in properties of finite-length polymers and proteins has grown rapidly within the past few years, not only because of the technological advances on the nanometer scale, but in particular due to the fact that the thermodynamics of small-scale systems is the key to the understanding of the biomechanical

principles of signal exchange and transport processes being relevant for life, as, for example, receptor–ligand binding between proteins or molecular flow through nanopores.

Since proteins play a key role in all biological processes, we will discuss their properties more specifically in the following.

1.2 Proteins

In order to get an impression of the complexity of the task of describing the relationship between chemical composition, geometric structure, and the function of individual macromolecules, we will now look more closely at one of the most prominent examples that represents an entire class of biomolecules: proteins.

1.2.1 The trinity of amino acid sequence, structure, and function

Proteins are highly specialized macromolecules performing essential functions in a biological system, such as control of transport processes, stabilization of the cell structure, and enzymatic catalyzation of chemical reactions; others act as molecular motors in the complex machinery of molecular synthetization processes. Chemically, proteins are built up of sequences of amino acid residues linked by peptide bonds. The polypeptide chain consists of a linear backbone with the amino-acid-specific side chains attached to it. The atomic composition of the protein backbone is shown in Fig. 1.1. Typical proteins consist of up to $N = 50, \ldots, 3000$ residues. The 20 different types of amino acids occurring in bioproteins are shown in Fig. 1.2. The side chains of these amino acids govern the specificity of each amino acid in protein folding processes and differ in chemical and physical properties under the influence of the surrounding solvent. Solubility in an aqueous environment is dependent on the occurrence of polar groups in the side chain, such as, e.g., the hydroxylic groups in serine and threonine. Hydrophobic side chains are insoluble and not very reactive in a polar environment. Typical large and strongly hydrophobic side chains such as, for example, phenylalanine or tryptophan possess aromatic rings. Others, like alanine or leucine with methine (-CH), methylene ($-CH_2$), or methyl ($-CH_3$) groups in the side chain, are aliphatic, i.e., these nonaromatic side chains only contain hydrogen and carbon atoms.

Primarily only arginine and lysine (positively charged) and aspartic and glutamic acid (negatively charged) contribute explicitly to the total charge of a protein in a neutral environment. In addition, histidine is typically positively charged in a slightly acidic environment, but neutral in neutral solution.

The frequency and the sequential arrangement of hydrophobic, polar, and charged amino acid residues in the amino acid sequence (also called the *primary structure*) of a protein are mainly responsible for the formation of a stable and unique native conformation. Protein conformations or segments of it are typically classified on different length scales. Parts of protein conformations that form local symmetric substructures are called *secondary structures*. These include helices, planar sheets (or strands), and turns. Substructures of this

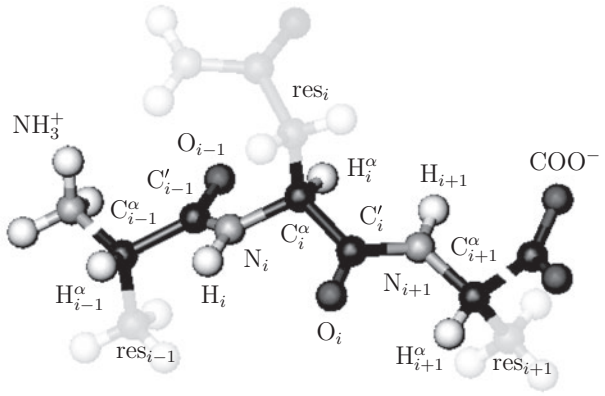

Fig. 1.1 Atomic composition of the protein backbone. Amino acids are connected by the peptide bond between C′ and N. Side chains or amino acid residues ("res") are usually connected to the backbone by a bond with the C^α atom (except proline, which has a second covalent bond to its backbone nitrogen).

type are common to all line-like objects and generally can be considered as the underlying geometry of linear polymers. The formation of secondary structures is not necessarily connected with the formation of hydrogen bonds, but these structures are essentially stabilized by hydrogen bonds. The whole conformation of a single protein, including aligned secondary structures, defines its *tertiary structure*. The tertiary fold typically consists of a very compact core of hydrophobic residues that is screened from the aqueous environment by a shell of polar amino acids being in direct contact with the solvent. This assembly of separate polar and hydrophobic parts is characteristic for proteins – it reduces entropy and thus ensures stability. Eventually, in large proteins or protein compounds, different hydrophobic domains can form widely independently. The global shape of the macromolecule that can include composites of individual tertiary domains is generally classified as *quaternary structure* [1–3].

The understanding of general aspects of the folding of proteins into native conformations is particularly essential, as the shape of the stable three-dimensional geometrical structure of a protein often determines the biological function of a protein – or its malfunction, if the protein has misfolded, refolded, or denatured.

The spectrum of protein functions in a biological organism is manifold. Proteins are involved in almost all cell processes. Ion channels and nanopores formed by membrane proteins control ion and water flow into and out of the cell [4]. Figure 1.3 shows the atomic structure of the membrane protein aquaporine, a pore embedded in the cell membrane that is permeable for water molecules. The efficiency is extreme. Up to 3 billion water molecules can rush through this pore per second [5, 6]. Other cell proteins like actin, for example, are responsible for the mechanical stability of the cell backbone. Actin polymerizes to filaments and these filaments can form stable networks. Besides cell stability, these networks also enable transport processes along these "tracks," e.g., vesicle transport mediated by myosin proteins. The interplay between actin and myosin is also important for the ability of muscle tissue to contract. Biochemical reactions are

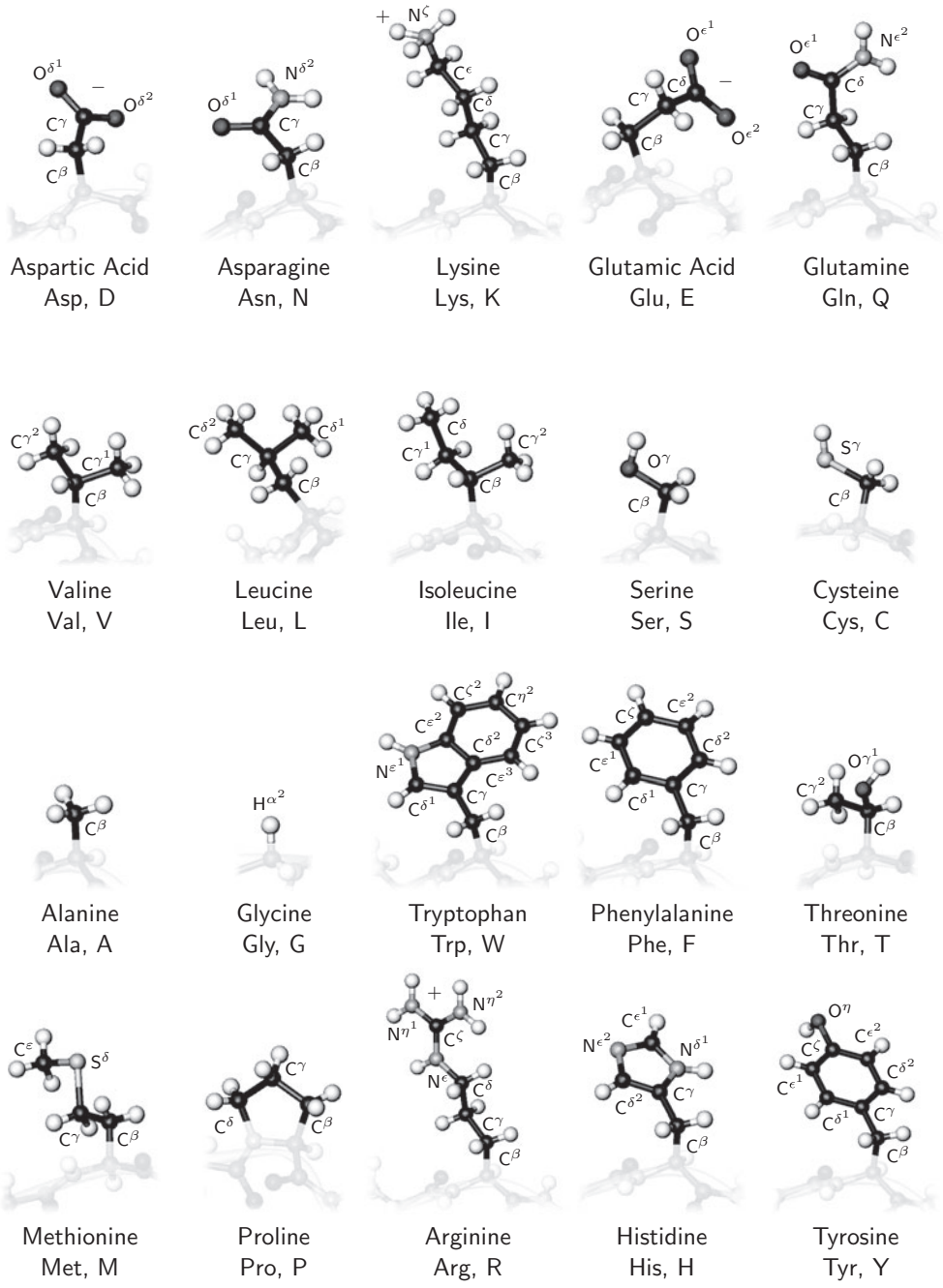

Fig. 1.2 The side chains of the 20 amino acids found in bioproteins. Side chains are bound to the backbone C^{α} atom, except proline, which has an extra bond to the backbone nitrogen N. Heavy atoms have been assigned standard labels [1]. There is no side chain for glycine, the free C^{α} bond is saturated by the hydrogen $H^{\alpha 2}$.

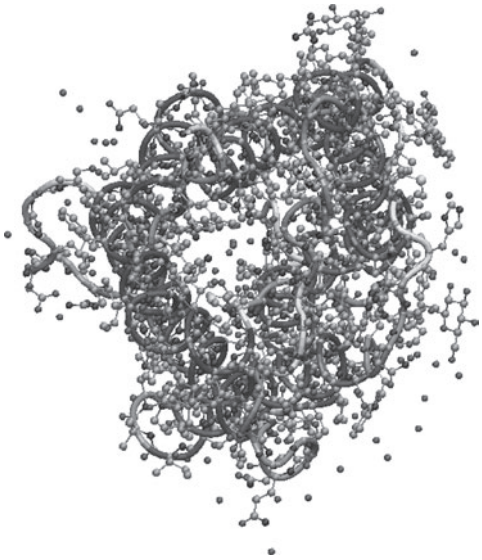

Fig. 1.3 Membrane protein *aquaporine*.

catalyzed by enzymes. General structural protein stability but also the ability to locally unfold and refold allow for receptor–ligand binding processes, which is necessary for enzymatic activity. Large proteins or protein compounds form complex nanoscale machines, which act as "molecular motors," as, for example, in the DNA/RNA polymerases and ATP synthase.

1.2.2 Ribosomal synthesis of proteins

The information about the amino acid composition of proteins is encoded in the DNA. In the polymerase II process, the genetic code is transcribed to the single-stranded messenger RNA (mRNA) sequence. The translation of the mRNA base code into amino acid sequences is part of the gene expression, and it is mediated by the ribosome. The ribosome itself is a macromolecular compound of large, multiple-domain proteins. A schematic snapshot of the ribosomal protein synthesis is shown in Fig. 1.4. Three successive bases along the mRNA strand always encode a single amino acid. The size of such a codon is intuitively clear: since there are four different bases building up the RNA code and the complementary base necessary to form a base pair is unique, a single-base codon could encode only 4 amino acids and 2 bases $4^2 = 16$ amino acid residues. Since 20 amino acids were identified in typical bioproteins, 3 bases are required to form a codon. The $4^3 = 64$ possibilities not only allow multiple, redundant codes for amino acids (which is a first kind of genetic error correction), but also enable the definition of start and stop codons, which are necessary to separate the codes for the different proteins within the linear RNA sequence [7, 8].

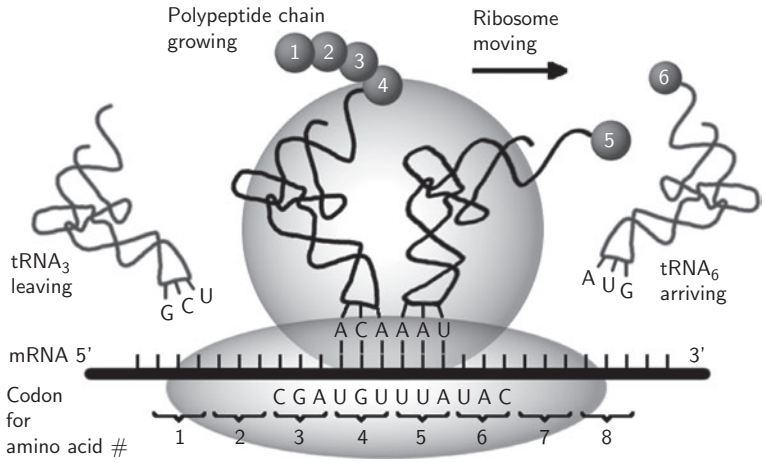

Fig. 1.4 The sequence of amino acids building up proteins is, based on the genetic DNA expression, synthesized in the ribosome.

After the ribosome has read a codon off the mRNA strand, a transfer RNA (tRNA) molecule connects to the mRNA codon. A tRNA molecule mainly contains a three-base section, the anticodon, which is complementary to a specific codon at the mRNA, and the associated amino acid residue. Thus the tRNA molecules serve as translators between the base codons and the amino acids. The ribosome separates the amino acid from the tRNA and attaches it to the already synthesized part of the protein sequence. This process continues until a stop codon is reached. Eventually, the protein is released into the aqueous solvent within the cell. It is widely believed that in this moment the protein is still unstructured and the formation of the functional structure is a spontaneous folding process. Larger proteins that would exhibit an increased tendency to misfold in the complex and crowded environment are often encapsulated in chaperons that assist in the folding process.

1.2.3 From sequence to function: The protein folding process

Anfinsen's refolding experiments [9] showed that the native conformation is *not* a result of the synthetization process in the ribosome. Rather, it is a dynamical process that strongly depends on intrinsic properties such as the amino acid sequence, but it is also influenced by the solvent properties and temperature of the surrounding solvent. Typical folding times are of the order of milliseconds to seconds. One of the most substantial problems in the understanding of protein folding is the "strategy" the protein follows in finding the unique conformation (geometrical structure), the so-called native fold. Thermodynamically, this native conformation represents the state of minimal free energy. Therefore, protein folding is not simply an energetic minimization process – it is also affected by entropic forces. This means that the folding trajectory is a stochastic process from a mostly random initial structure toward the global free-energy minimum, thereby circumventing free-energy barriers. The free-energy landscape of a protein, considered as a function of the protein's

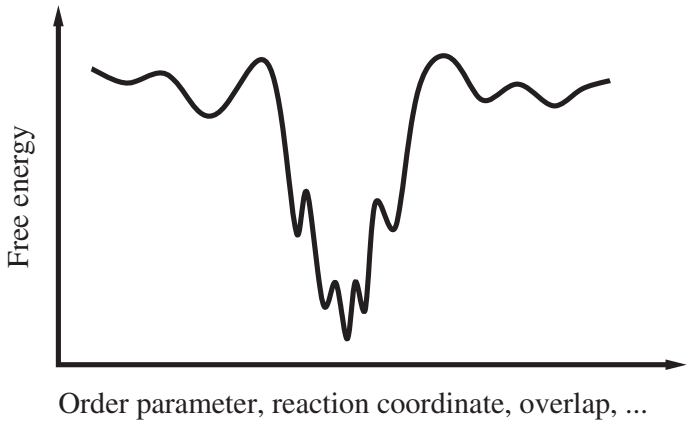

Fig. 1.5 Sketch of a free-energy landscape for a protein under folding conditions as a function of a single cooperativity parameter, which is often referred to as a reaction coordinate in analogy to chemical reaction kinetics, as an order parameter by considering the folding process as an analog of a thermodynamic phase transition, or as a kind of overlap parameter similar to what is often used in descriptions of metastable systems such as spin glasses.

degrees of freedom, is commonly assumed to be extremely rugged and complex and it seems paradoxical that the protein is able to find the "needle-in-a-haystack" structure by a stochastic search process within a relatively short time period.

Protein folding is, however, also a process of high cooperativity, i.e., structure formation requires a collective arrangement of at least a subset of the degrees of freedom. For this reason, it is expected that a single or a few cooperativity parameter(s) – comparable to order parameters in thermodynamic phase transitions – allow(s) for the discrimination between dominating macrostates, i.e., the structural "phases." A strongly simplified sketch of such a free-energy landscape as a function of a single cooperativity parameter is shown in Fig. 1.5. For stability reasons, the single funnel-like global free-energy valley is sufficiently deep to prevent thermal unfolding. Local folding processes can cause weakly stable or metastable conformations that slow down the folding process. Thus, the folding channel is not necessarily smooth and local free-energy minima can be present in the funnel.

The assumption of the existence of a reduced set of relevant collective degrees of freedom thus enables a generalized view of folding processes as structural or conformational transitions and their classification. Indeed, there has been enormous progress in this direction within the past years, and candidates with comparatively simple folding trajectories were identified. This regards, e.g., single-exponential or downhill folding, where no barriers hinder and slow down the folding process. Another prominent example is two-state folding, a "first-order-like," "discontinuous" transition with a single barrier. More complex are "folding-through-intermediates" events with more than one free-energy barrier or even metastability, where the native fold is degenerate and, therefore, the formation of different structures is (almost) likely probable. The latter case is obviously important for proteins involved in mechanical or motoric processes, where local refolding can be necessary for fulfilling a specific biological function.

1.3 Molecular modeling

Structure formation at the atomic level is, in principle, a traditional quantum-chemical many-body problem. Amino acids occurring in bioproteins contain between 7 (glycine) and 24 (tryptophan) atoms.[6] Thus, typical bioproteins consist of hundreds to tens of thousands of atoms. In general, the structural properties of macromolecules depend on two classes of chemical bonds: *covalent* and *noncovalent* bonds. Covalent bonds are based on common electron pairs shared between atoms and stabilize the chemical composition of the molecule. On the other hand, noncovalent bonds[7] are based on much weaker effective interactions due to screening, polarization effects, or dipole moments, partly induced by the surrounding polar solvent. These interactions are responsible for the three-dimensional structure of macromolecules in solvent.

1.3.1 Covalent bonds

The formation of a covalent bond between atoms is a pure quantum-mechanical effect and due to an effective pairwise attraction between electrons in outer, unsaturated shells of two atoms. This spin-dependent exchange interaction overcompensates the electrostatic repulsion between the electrons and results in an electron pair, which is shared by the atoms involved. Covalent bonds are very stable and a thermal decomposition, e.g., at room temperature, is extremely unlikely. The dissolution energy of biochemically relevant covalent bonds lies between 50 kcal/mole (disulfide bridges S–S) and 170 kcal/mole (C = O double bonds). For comparison, the thermal energy at room temperature $T_r = 300$ K is $RT_r \approx 0.6$ kcal/mole.[8] This energy is also not sufficient to excite vibrations of covalent bonds at room temperature. Therefore, the effective bond lengths, i.e., the distances between the atom cores, are rigid. Furthermore, the bond angles between two successive covalent bonds are relatively rigid. In any event, weak vibrational fluctuations are thermally excitable, but the typical fluctuation widths are comparatively small and usually do not exceed 5°. In proteins, covalent bonds obviously stabilize the sequence of the amino acids linked by covalent peptide bonds, i.e., its primary structure. Although torsional degrees of freedom also are affected by covalent bonds, a subset of torsional angles is widely flexible: the so-called *dihedral torsion angles*. Figure 1.6 shows a conformation of a small peptide, with the end-groups NH_3^+ and COO^- and the amino acid phenylalanine (Phe) highlighted. The Ramachandran dihedral angles in the backbone are typically denoted ϕ (torsional angle between atoms C'_{i-1}, N_i, C_i^α, and C'_i of the $(i-1)$th and ith amino acid) and ψ (between N_i, C_i^α, C'_i, N_{i+1}). The angles ϕ and ψ are comparatively flexible in the interval $-180° < \phi, \psi \leq 180°$, because the torsional barriers imposed by the electronic properties

[6] Numbers of atoms refer to uncharged amino acids within a polypeptide chain.
[7] Typically associated with *nonbonded interactions*.
[8] The gas constant in molar units is $R = N_A k_B \approx 1.99 \times 10^{-3}$ kcal/K mole, where $N_A \approx 6.02 \times 10^{23}$ mole^{-1} is the Avogadro constant and $k_B \approx 3.30 \times 10^{-27}$ kcal/K is the Boltzmann constant.

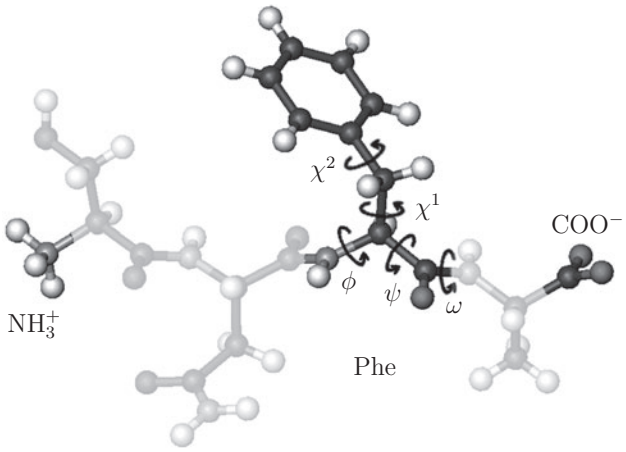

Fig. 1.6 Definition of the backbone dihedral angles ϕ, ψ, and ω. Exemplified for phenylalanine, also the only two side-chain degrees of freedom χ^1 and χ^2 are denoted. The convention is that the torsional angles can have values between $-180°$ and $+180°$, counted from the N-terminus (NH_3^+) to the C-terminus (COO^-) according to the right-hand rule and in the side chains starting from the C^α atom.

of the covalent bonds are rather weak. An exception is proline, where the particular geometry of the side chain restricts ϕ to a value close to $-75°$. The angle ω is associated with the torsion of the peptide bond C_i^α, C_i', N_{i+1}, and C_{i+1}^α. However, the sp2 hybridizations of the C′ and N valence electrons and a p-electron uninvolved in the hybridizations that form an electron cloud surrounding the C′-N peptide bond entail a large torsional barrier. Thus, $\omega \approx 180°$, and C_i^α, C_i', N_{i+1}, and C_{i+1}^α form an almost planar *trans* conformation. Proline is special as it is bound to three massive radicals instead of the usual two. For this reason, there is also a non-negligible amount (about 10%) of proline involving peptide bonds in *cis* conformation ($\omega \approx 0°$) [2]. Phenylalanine possesses two torsional side-chain angles, χ^1 and χ^2, that can be thermally activated under physiological conditions. Depending on the type of amino acid and its atomic composition, the number of torsional side-chain angles varies.

Thus, the three-dimensional geometric structure of proteins is little dependent on covalent bonds, and it is rather due to the much weaker effects between nonbonded atoms. Nonetheless, the rigidity of covalent bond lengths and bond angles affects the process of structure formation (e.g., the folding of a protein into the native state with lowest free energy). Generally, the steric constraints promote frustration and metastability and thus the existence of stable native conformations is a very particular property of the comparatively small number of functional bioproteins selected by evolution. However, the majority of all possible amino acid sequences suffers from degeneration effects, e.g., nonfunctional metastability, and also from only weakly stable conformations under physiological conditions. For this reason, such sequences play only a very minor role in biological systems (molecular motors, for example, require only small activation barriers for motion, but refolding often affects only small parts of the structure).